Statistics and Computing (1993) **3**, 27–35

The stalactite plot for the detection of multivariate outliers

A. C. ATKINSON and H.-M. MULIRA

Department of Statistics, London School of Economics, London WC2A 2AE, UK

Received January 1992 and accepted September 1992

Detection of multiple outliers in multivariate data using Mahalanobis distances requires robust estimates of the means and covariance of the data. We obtain this by sequential construction of an outlier free subset of the data, starting from a small random subset. The stalactite plot provides a cogent summary of suspected outliers as the subset size increases. The dependence on subset size can be virtually removed by a simulation-based normalization. Combined with probability plots and resampling procedures, the stalactite plot, particularly in its normalized form, leads to identification of multivariate outliers, even in the presence of appreciable masking.

Keywords: elemental set, Mahalanobis distance, masking, Monte Carlo testing, normalized distance, outliers, simulation, stalactite plot

1. Introduction

The presence of one or two outliers in multivariate data can often be detected by calculation of the Mahalanobis distance for each observation. However, the distances may fail to indicate the presence of any outliers when many are present. This masking effect is due to the influence of the outliers on the estimates of the means and covariance matrix used in calculating the distances. To overcome this problem Rousseeuw and van Zomeren (1990) suggest the use of distances based on robust estimates of location and covariance. However, Cook and Hawkins (1990) show that this procedure may indicate a plethora of outliers, the identity of which can change dramatically with small changes in the parameters of the algorithm for robust estimation. They commend instead backward procedures in which outliers are sequentially detected and deleted, starting from all n observations. It is the purpose of this paper to consider rather a forward procedure which starts by using a small random subset of the data for estimation of the means and covariances required for calculation of the Mahalanobis distances. The size of the subset is then increased in such a way as to exclude outliers. The procedure unambiguously identifies the outliers in the example studied by Cook and Hawkins. It also has the advantage of computational modesty while yielding a simple graphical summary, the stalactite plot.

We start in Section 2 with a discussion of deletion Mahalanobis distances. In regression, 'leave-one-out' diagnostics provide valuable information on outliers and influential observations. We show that similar information is not obtained by use of leave-one-out Mahalanobis distances. Section 3 describes the forward algorithm for the selection of observations to be used in the sequential construction of Mahalanobis distances. The crucial idea is that, given distances using m observations for estimation of the means and covariances, the $m+1$ observations to be used for the next set of distances are chosen to be those with the $m+1$ smallest distances. Thus an observation can be included in the estimates for some values of m, but can later be excluded as m increases. Two examples are presented in Section 4: use of the stalactite plot graphically illustrates the evolution of the set of outliers as the size of the fitted subset m increases. Probability plots are used to provide a distributional framework for interpretation of the distances.

When the subset size m is small relative to the number of observations, the stalactite plot typically indicates the presence of many outliers, the number decreasing as m increases. This effect of the size of m, compounded by the way in which successive subsets are chosen, is overcome by a plot of normalized distances introduced in Section 5. The normalization is obtained from stalactite analyses of simulated data and depends on the dimension of the data.

If there are many outliers and the data are sparse, a single run of the stalactite analysis may fail to detect the outliers. Section 6 illustrates the properties of repeated analyses with different random initial subsets. The procedure, combined with the normalization of Section 5, leads to the clear identification of the four outliers in the data example used by Cook and Hawkins, the modified wood gravity data of Rousseeuw and Leroy (1987, p. 243). Cook and Hawkins state the need for 'better graphical displays' of multivariate outliers. We believe that the stalactite plot, and other plots in this paper, provide such displays.

2. The deletion Mahalanobis distance

Let y_k^{T} be the kth of n observations on a p-variate normal population. Then the squared Mahalanobis distance for observation k is

$$d_k^2 = (y_k - \bar{y})^{\mathrm{T}} S^{-1} (y_k - \bar{y}) = r_k^{\mathrm{T}} S^{-1} r_k, \qquad (1)$$

where $\bar{y}$ is the $p \times 1$ vector of means and S is the sample covariance matrix with

$$S_{ij} = \sum_k (y_{ki} - \bar{y}_{.i})(y_{kj} - \bar{y}_{.j})/(n-1) = \sum_k r_{ki} r_{kj}/(n-1). \qquad (2)$$

Asymptotically the d_k^2 follow a chi-squared distribution on p degrees of freedom. If $\bar{y}$ and S were not estimated but were known population parameters, outlying values of y_k would yield large values of the squared distance d_k^2. However, the effect of such values on the estimation of $\bar{y}$ and S leads to the rapid breakdown of the Mahalanobis distance for the detection of outliers, particularly if several outliers are present.

This problem is not overcome by consideration of the deletion Mahalanobis distance

$$d_{(k)}^2 = \{y_k - \bar{y}_{(k)}\}^{\mathrm{T}} S_{(k)}^{-1} \{y_k - \bar{y}_{(k)}\}, \qquad (3)$$

where $\bar{y}_{(k)}$ and $S_{(k)}^{-1}$ are the estimated means and covariances with observation k deleted. It is straightforward that

$$y_k - \bar{y}_{(k)} = \frac{n}{n-1}(y_k - \bar{y}) = \frac{n}{n-1} r_k. \qquad (4)$$

Further relationships are most easily found from the standard deletion formulae of least squares regression diagnostics (for example, Cook and Weisberg, 1982, §2.2; Atkinson, 1985, §2.2). The change in the elements of S on deletion follows from the change in the residual sum of squares of a regression model on deleton of an observation, which yields

$$(n-2) S_{ij(k)} = (n-1) S_{ij} - n r_{ik} r_{jk}/(n-1). \qquad (5)$$

To find $S_{(k)}^{-1}$, (5) is applied to the formula for updating a matrix inverse (Cook and Weisberg, 1982, p. 210; Atkinson, 1985, p. 19). The combination of this form of the inverse with (4) leads to the compact expression for the squared deletion distance

$$d_{(k)}^2 = \frac{(n-2)n^2}{(n-1)^3}\{d_k^2/(1 - n d_k^2/(n-1)^2)\}. \qquad (6)$$

Thus the deletion distance is a monotone function of the usual Mahalanobis distance and so provides no additional diagnostic information.

In regression models deletion of single observations provides extra information because of the presence in the deletion formulae of the leverage measure h_i as well as of the residuals r_i. But in the calculation of the distances of this section there is no effect of leverage: since we are only concerned with means, all $h_i = 1/n$. Thus the simple reduction produced by the formulae of this section would not apply if distances were being calculated using the residuals from multivariate regression.

3. The forward identification of outliers using Mahalanobis distance

The results of the previous section show that there is no diagnostic information in single deletion distances which can be added to the information about outliers obtained from the standard distance d_k^2 given by (1). We therefore consider instead distances for all n observations calculated using outlier free estimates of the means and covariances. Rousseeuw and van Zomeren achieve this by the use of robust estimates based on the minimum volume ellipsoid covering half the data. We instead use the standard estimates of the previous section, but from a subset of m observations chosen to be unlikely to contain outliers.

Suppose that $m < n$ observations have been used to calculate the means $\bar{y}$ and the covariance matrix S. Using these estimates n Mahalanobis distances can be calculated. If the m observations used in fitting are all outlier free, any outliers present will give rise to large Mahalanobis distances. The method for forward identification of outliers which is the basis of this paper uses the $m + s$ observations with smallest distances to calculate new estimates of the mean and covariances. Usually $s = 1$. Provided the outliers were correctly identified by the fit using m observations, this fit will also exclude outliers, since they give rise to large distances. Outliers will only be included as m approaches n, when no good observations remain to be introduced into the fit. Hadi (1992) uses the same forward algorithm starting from robust estimates of the means and covariances for calculation of the initial Mahalanobis distances. His forward search terminates when m is the median of the number of observations when allowance is made for the effect of fitting. The method used here continues until $m = n$.

In the examples we start with a randomly selected subsample of observations with $m = p + 1$, the smallest number from which the distances can be calculated. This starting point is the same as that used by Rousseeuw and van Zomeren in their random search algorithm for the minimum volume ellipsoid. It is also close to that used in the elemental set algorithm for approximate least median of squares regression (Rousseeuw, 1984) and for the identification of multiple outliers in regression (Hawkins *et al.*, 1984) where, however, elemental sets of p observations are used. In the examples we take $s = 1$, so that one more observation is included at each step. For larger data sets than those considered here, larger values of s would be appropriate. As with the elemental set algorithm, several starting points can be used to increase the probability of obtaining initial distances based on an outlier-free set. This technique is illustrated in Section 6. However this does not usually seem to be necessary. The repeated use, as m increases, of the observations yielding the m smallest Mahalanobis distances often results in the exclusion of outliers, even if they are included at an early stage of the algorithm. This point is illustrated in Example 2 of the next section.

The result of this analysis is a matrix of $(n - p - 1) \times n$ Mahalanobis distances. As we shall see, the pattern may be quite irregular when $m < n/2$, although the normalized distances of Section 5 produce more regular patterns, at the cost of more computing. For larger values of m the pattern of distances settles down, giving large distances for the outliers, until m is so large that outliers begin to be included in the subset used for parameter estimation. As this happens, masking begins to occur until, when $m = n$, there may be no apparent outliers.

4. The stalactite plot

An example of a stalactite plot is given in Fig. 1(a). The subsample size used in estimation is m. The plot indicates, for increasing m downwards, those observations for which the Mahalanobis distance is sufficiently large for the observation to be considered an outlier. The cut-off point used to define an outlier is the maximum expected value from a sample of n chi-squared random variables on p degrees of freedom, approximated by

$$E(\max \chi_p^2) = \chi_p^2\{(n - 0.5)/n\}, \qquad (7)$$

the actual approximation not being important.

The stalactite plot shows how the pattern of suspected outliers changes with m. It is also informative to look at the values of the Mahalanobis distances, typically when m is 80% or 90% of the sample size n. These values can be presented as index plots. However, probability plots are more helpful in interpreting the magnitudes of the distances for suspected outliers.

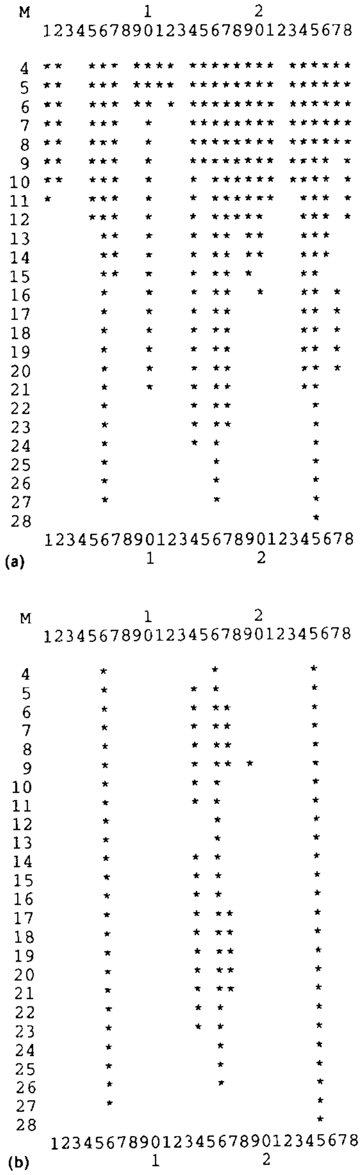

Fig. 1. *Example 1: brain and body weight. Stalactite plot: (a) original distances; (b) normalized distances. Subsample size M plotted against observation number*

The remainder of this section is devoted to the presentation of examples. In order to provide a comparison of our findings with those of Rousseeuw and van Zomeren and of Cook and Hawkins, some of their examples are reanalysed.

Example 1. Brain and body weight

Rousseeuw and van Zomeren motivate their paper with an analysis of the logarithms of brain and body weight of 28 animals given by Rousseeuw and Leroy (1987, p. 57). These data are an augmentation of part of a set given by Weisberg (1985, pp. 144–5) to illustrate the importance of transformations in the analysis of data. A bivariate scatter plot of the logarithms (to base 10) of the data is given by Rousseeuw and van Zomeren. This plot, and their analysis, indicate that observations 25, 6 and 16 (the three dinosaurs added to Weisberg's data on mammals) are the most extreme outliers. In addition observations 14 and 17 (human beings and rhesus monkeys) may not agree with the bulk of the data.

Figure 1(a) shows a stalactite plot for these data. Since $p = 2$, the first set of distances are calculated for a subset of $m = 3$ randomly chosen observations. The first set chosen by the forwards procedure therefore has $m = 4$. Initially nearly all the distances are large enough to indicate outliers but, as m increases, there is a gratifyingly smooth evolution towards a few persistent outliers. For $m > 21$ these are the five observations already mentioned, namely 6, 14, 16, 17 and 25. For $m > 24$ only three outliers are indicated, 6, 16 and 25. The masking effect of these outliers is shown by the distances for the full sample: only observation 25 has a distance greater than the maximum expected chi-squared.

This analysis is supported by two further sets of figures. The index plots of Fig. 2 show the Mahalanobis distances when $m = 90\%$ of n, that is 25, and when all observations are used to estimate the means, variances and covariances. In Fig. 2(a) observations 6, 16 and 25 all have large values, with that for observation 14 only just below the cutoff. At $m = 80\%$ of n the plot, not shown here, is very similar, again with three clear outliers but with two further distances just greater than the cut-off value. The plot when all the data are used in fitting, i.e. $m = n$, Fig. 2(b), again clearly shows the effect of masking — only observation 26 is indicated as outlying, and that only just. In this example the masking effect arises because observations 6, 16 and 25 form a distinct cluster, the presence of which distorts the estimates of the means and covariance matrix in such a way as to reduce the distances for these observations.

In the second analysis of the distances we provide a sampling reference by using chi-squared probability plots. The plots of the squared distances for 80% and 90% subsamples are very similar, showing three clear outliers — the smaller values for observations 14 and 17 hardly lie off the probability plot. Only that for the 90% subsample is given here as Fig. 3(a). As Fig. 3(b) shows, when all the observations are used in fitting, there is no evidence of the presence of outliers, the plot sensibly following a straight line. The final conclusion, that there are three gross outliers, agrees with inspection of Rousseeuw and van Zomeren's Fig. 1.

Example 2. Synthetic data of Hawkins et al. (1984)

A second example used by Rousseeuw and van Zomeren is the three explanatory variables of an artificial data set generated by Hawkins *et al.* (1984) to illustrate the effect of outliers at leverage points on least squares estimates of

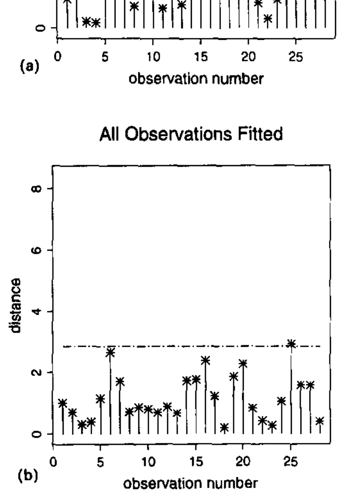

Fig. 2. *Example 1: brain and body weight. Index plots of Mahalanobis distances: (a) $m = 25$; (b) $m = n$*

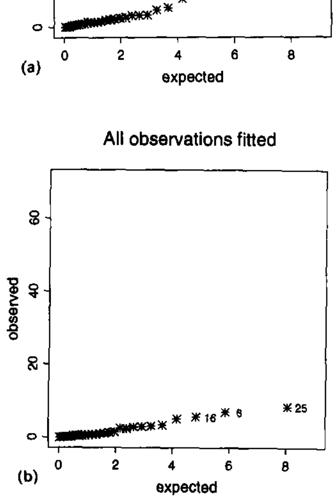

Fig. 3. *Example 1: brain and body weight. Chi-squared probability plots of squared Mahalanobis distances: (a) $m = 25$; (b) $m = n$*

regression coefficients. For these data $n = 75$, $p = 3$ and there are two groups of outliers: observations 1–10 and 11–14.

Figure 4(a) is a stalactite plot of the Mahalanobis distances for this example, arbitrarily shortened by the omission of low even values of m. For $m > 49$, only observations 1–14 are indicated as outliers, a pattern which persists until m is so large that outliers start to be included in the subset used for estimation. Index plots and probability plots similar to Figs 2 and 3 support this identification of outliers in a manner parallel to that of the previous example, so are not repeated here, but there is one new point.

The sequential procedure yielding the stalactite plot depends, in theory, on the initial subsample being clear of outliers, so that unbiased estimates are obtained of means and covariances for calculating the distances. Since the initial subsample is selected at random, one or more outliers could well be included, perhaps leading to a subsequent failure to identify some of the outliers. To investigate this effect, the stalactite plot for the data of Hawkins *et al.* was recalculated using observations 1–14, that is all the outliers, as the initial subsample. The plot at first leads to identification of nearly all observations, other than 1–14, as outliers, followed, for increasing m, by a reduction in the number of apparent outliers. But from $m \geq 48$, observations 1–14 are identified as the only outliers, as they are in Fig. 4(a). The plot is not given here as the point is made more forcibly by plotting the normalized distances introduced in the next section.

5. The normalized stalactite plot

The stalactite plots of Figs 1(a) and 4(a) allow identification of the outlying observations. However, particularly when m is small, an appreciable number of outliers are identified. In this section we describe a normalization of the plot which, at the cost of extra computation, greatly reduces the number of apparent outliers for small m and leads to clearer identification of the observations which are truly outlying.

Let the squared Mahalanobis distance (1) when all n observations are used in calculating $\bar{y}$ and S be denoted $d_k^2(n)$, with $d_k^2(m)$ the distance when a subset of size m is used. Then if

$$T(m) = \sum_{k=1}^{n} d_k^2(m), \tag{8}$$

it is well known, and follows from (1) and (2), that $T(n) = p(n-1)$. In general $T(m)$ will be greater than $T(n)$. Furthermore, the way in which successive subsets are selected by the forward method of Section 3 means that, particularly for small m, $T(m)$ will be very much greater than $p(n-1)$. The consequent identification of a

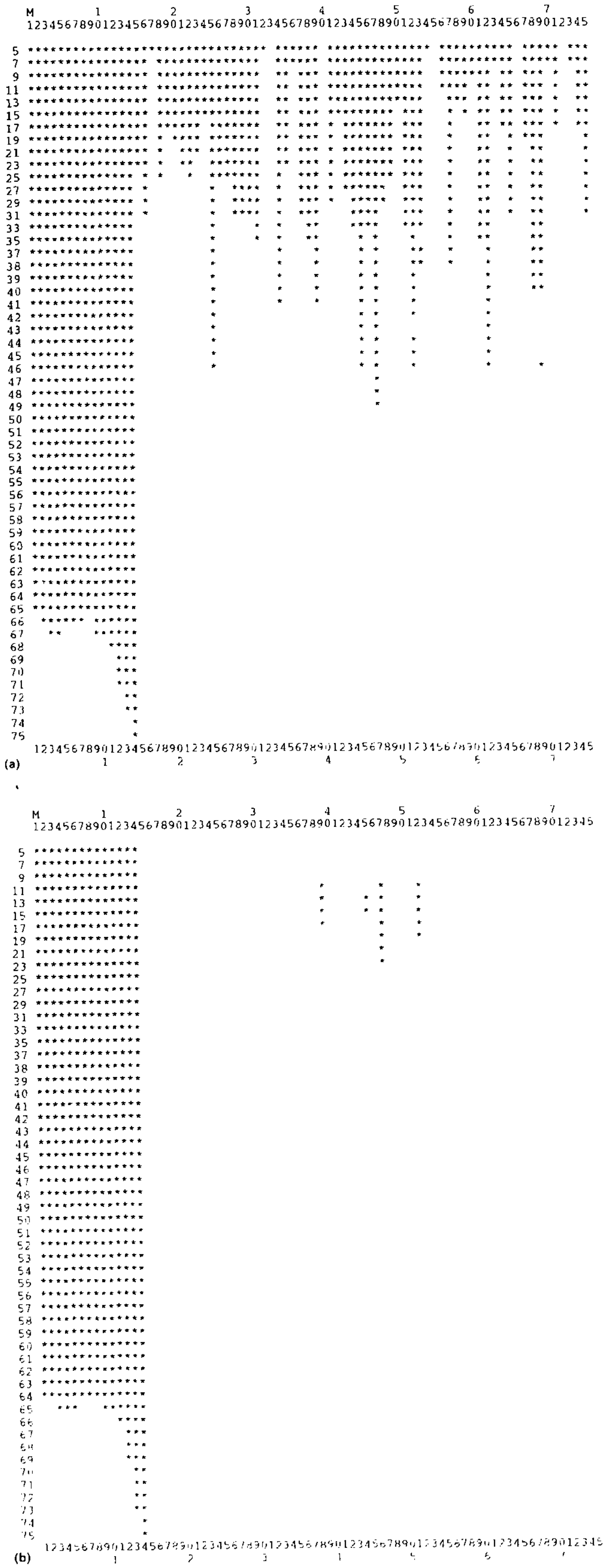

Fig. 4. *Example 2: synthetic data of Hawkins et al. Stalactite plot: (a) original distance; (b) normalized distances. Subsample size M against observation number*

large number of outliers for small m can be corrected by a simulation based normalization.

To normalize the distances N stalactite analyses are performed, each on a different simulation of multivariate normal data with dimensions p and n the same as those of the original data. Because the distances (1) are invariant to the population mean and variance, the simulated data can be merely pn independent standard normal random variables. For each value of m from $p + 2$ to n the squared distances are summed and then averaged over the N simulations, yielding a series of averaged totals $\bar{T}(m)$. The normalized squared distances are then

$$\tilde{d}_k^2(m) = p(n-1)\, d_k^2(m)/\bar{T}(m) \qquad (9)$$

We have found that taking $N = 100$ gives stable estimates of the normalized distances. Smaller values also appear acceptable but, for problems the size of those in this paper, the computations are not so extensive that we have felt it necessary to investigate this point.

The results of the normalization are astounding. Fig. 1(b) is the normalized version of the stalactite plot of Fig. 1(a) for the brain and body weight data. For all m from 4 to 26 the three dinosaurs are revealed as outlying. Over much of this range observations 14 and 17 are also identified by the plot, which only once identifies any other observation as outlying.

The results are equally incisive for the data of Hawkins *et al.* Fig. 4(b) shows observations 1–14 as outlying for $m = 5$ to 64. Only briefly are a few other observations shown on the plot, and none for $m > 23$. If the procedure is started with only outliers in the initial subset, the plot of Fig. 5 shows how the search recovers, without indicating any other outliers on the way.

In order to reduce the size of the stalactite plots of Figs 4 and 5, alternative lines have been omitted for low values of m. For large n, when the structure changes slowly, only one in every k lines need be plotted, even if the increment s in the calculations equals 1. The examples in this section show that the structure does indeed change slowly for the normalized plot, with many observations never figuring as outliers. If it is required to compress the stalactite plot further, these columns could be entirely omitted.

The only difference between the normalized and unnormalized stalactite plots is that the distances have been reduced by the use of (9). An alternative way of describing this is that the critical point for appearing on the stalactite plot (7) has been multiplied by $\bar{T}(m)/\{p(n-1)\}$. In this framework the normalized stalactite plot is a visualization of a Monte-Carlo testing procedure, the value of $\bar{T}(m)$ being obtained by a simulation experiment.

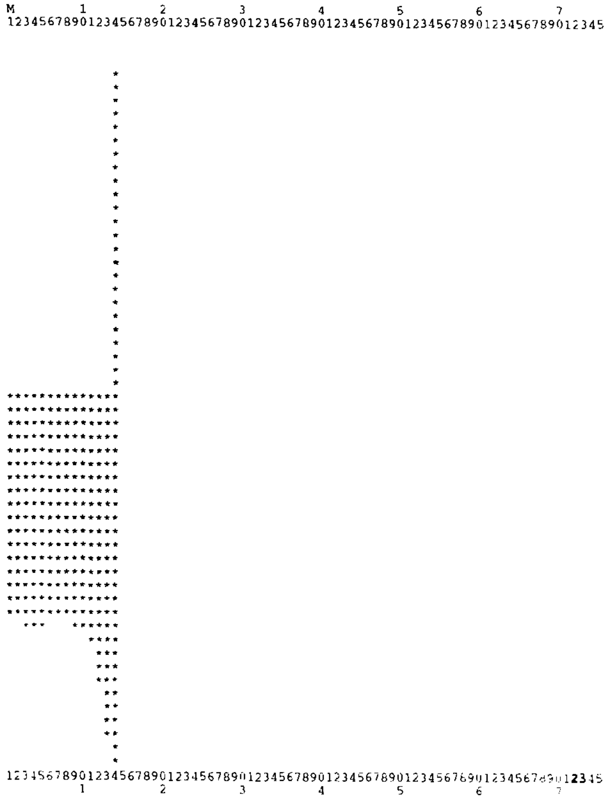

Fig. 5. *Example 2: synthetic data of Hawkins et al. Stalactite plot of normalized distances with initial subset containing only outliers. Subsample size M plotted against observation number*

6. Repeated sampling

In Example 2 forward calculation of the Mahalanobis distances was able to recover from an initial subset which contained all the outliers. But the method cannot always do this, particularly when there are many outliers or the data are sparse. One measure of sparseness is the ratio n/p which is respectively 14 and 25 for Examples 1 and 2. However, in the next example this ratio is 4 and the pattern of apparent outliers does indeed depend on the starting point of the algorithm. This dependence is removed by multiple repetition of our forward selection method combined with investigation of the normalized stalactite plots yielding the largest maximum distances. The result is again an unambiguous identification of the outliers.

Example 3. Modified wood gravity data

Cook and Hawkins illustrate their claim that Rousseeuw and van Zomeren's procedure can produce 'outliers everywhere' by an analysis of the wood gravity data (Draper and Smith, 1966, p. 227) modified by Rousseeuw and Leroy (1987, p. 259). This is again the set of explanatory variables from a multiple regression problem, but now $p = 5$ and $n = 20$. The data are thus in a higher-dimensional space than are the previous examples ($p = 2$ and 3) and are also fewer, so problems due to sparseness may be anticipated.

Cook and Hawkins find a wide variety of 'outliers' depending upon the number of observations used to calcu-

```
 M          1         2
   12345678901234567890

 7 * ******  * * **  *
 8 * ******  * * **  *
 9 * ******  *   **  *
10   ******  *   **  *
11   ** ***  *    *  *
12   ** ***  *    *  *
13    * ***  *    *  *
14    * * *  *    *  *
15    * * *  *       *
16    * * *          *
17      * *          *
18
19
20
   12345678901234567890
(a)         1         2
```

```
 M          1         2
   12345678901234567890

 7
 8    * *            *
 9    * * *          *
10    * * *          *
11    * * *          *
12    * * *          *
13    * * *  *       *
14    * * *          *
15    * * *          *
16    * * *          *
17
18
19
20
   12345678901234567890
(b)         1         2
```

Fig. 6. *Example 3: modified wood gravity data. Stalactite plot showing observations 4, 6, 8 and 19 as outliers: (a) original distances; (b) normalized distances. Subsample size M plotted against observation number*

late the robust estimates of means and covariances. We were more fortunate. The stalactite plots of Fig. 6 from our first analysis of the data, in particular the normalized stalactite plot of Fig. 6(b), clearly show observations 4, 6, 8 and 19 as outliers, which are the four observations modified by Rousseeuw and Leroy. This result is confirmed by the probability plot of Fig. 7(a) for $m = 16$, that is 80% of n.

That this result was obtained from our first analysis was something of a fluke. Table 1 summarizes the results of 100 repetitions of the stalactite analysis at the point when $m = 16$. The stalactites are ordered by the maximum value of $d_k^2(16)$, which is 115.0 for the most extreme solution, yielding the plots of Fig. 6 and Fig. 7(a), for which the outliers are observations 4, 6, 8 and 19. This solution was obtained 15 times. For the next most extreme solution the maximum squared distance is 61.0 and the normalized distances indicate that observations 7 and 9 are outlying, a conclusion supported by the probability plot of Fig. 7(b). This solution occurred only once, although 24 other stalactites also indicate the same pair of outliers: the differences in values of maximum $d_k^2(16)$ are caused by varying choices of the two observations, other than 7 and 9, which are excluded from the fit.

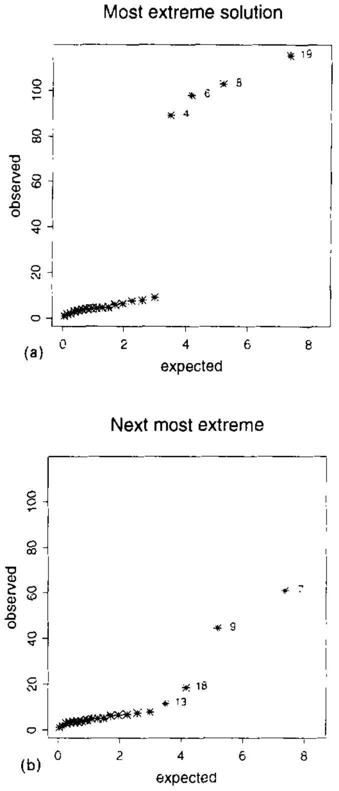

Fig. 7. *Example 3: modified wood gravity data. Chi-squared probability plot of squared Mahalanobis distances for m = 16: (a) outliers from Fig. 6; (b) second most extreme solution in Table 1*

The presence of several possible sets of outliers seems an inescapable feature of the analysis of sparse data. Viewed as an optimization, maximization of the maximum squared distance for a specified m is a problem with several local maxima. As is standard in optimization practice, multiple random starts have been employed, leading to the results summarized in Table 1. In the analysis of data, rather than the analysis of outlier detection techniques, one occurrence of an extreme solution is sufficient to indicate the presence of masked outliers.

7. Discussion

Forward calculation of Mahalanobis distances by the resampling method of this paper provides an alternative to the methods of Rousseeuw and van Zomeren for the identification of multivariate outliers. The stalactite plot, especially in its standardized form, together with its associated index and probability plots, has been shown to provide clear conclusions for two examples in which the observational points are not too sparse. For both these examples 100 replicates of the stalactite analysis all led to the identi-

Table 1. *Modified wood gravity data. Results of 100 stalactites analysed when 16 observations are fitted. Outlying normalized distances*

	Observation Number																					
Maximum Squared	0									1										2	Number of	
Distance	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	Outliers	Occurrences
115.0				#		#		#											#		4	15
61.0							#		#												2	1
50.7							#		#												2	2
49.7							#		#												2	2
49.2							#		#												2	9
46.9										#			#							#	3	8
42.8										#						#					2	2
40.2							#														1	1
37.8							#														1	6
37.1							#		#												2	11
36.1							#														1	1
35.5							#														1	1
35.2																#					1	3
32.2							#														1	16
29.3										#											1	1
29.0																#					1	1
28.6												#									1	1
26.2												#									1	3
25.5																					0	1
25.4																					0	1
24.9																					0	3
24.2																					0	1
23.7																					0	2
22.7																					0	1
22.4																					0	1
21.3																					0	1
20.2																					0	3
19.1																					0	1
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0		
	0									1										2		

fication of outliers given in Sections 4 and 5. Calculation of the index and probability plots at several subsample sizes provides a method of investigating the sensitivity of inferences to the assumed outlier level. This is analogous to the coverage parameter of Rousseeuw and van Zomeren, investigated by Cook and Hawkins, and to the assumed level of contamination in the derivation of an optimal M-estimate for location.

The computational requirements of our method are modest when compared with the effort needed to find the minimum-volume ellipsoid. An extensive comparison of algorithms for this problem is made by Woodruff and Rocke (1992), who show that the random search algorithm cannot be efficient. The increased precision yielded by the normalized plot of Section 5 requires increased computaton which is however still small compared with that of methods for finding the minimum volume ellipsoid. Whatever method is used for the detection of multivariate outliers, sparseness is potentially a problem. The repeated sampling procedure of Section 6 requires a similar amount of computing to that needed for the normalized stalactite plot. We have shown it provides a reliable method of detecting multivariate outliers where methods described in the references have failed.

Note

Identification of the most extreme sets of outliers yielded by the repeated sampling method of Section 6 is aided by calculation for each run of the volume of the smallest ellipsoid containing half the data. Small values of this volume result from outlier free estimates of the parameters and yield extreme distances. A simple ordering of the searches is thus possible. This idea is explored in Atkinson (1993).

Acknowledgements

We are grateful to the referees whose comments inspired us to the work leading to the normalized stalactite plot, as well as for comments which, we trust, have led to improved clarity of presentation. The computations for this research were performed on equipment funded, in part, by the Science and Engineering Research Council of the United Kingdom under its Complex Stochastic Systems Initiative.

References

Atkinson, A. C. (1985) *Plots, Transformation, and Regression.* Clarendon Press, Oxford.

Atkinson, A. C. (1993). Robust estimation for outlier detection. In *Data Analysis and Robustness*, S. Morgenthaler and W. Stahel (eds.) Birkhäuser, Basel.

Cook, R. D. and Hawkins, D. M. (1990) Comment on Rousseeuw and van Zomeren (1990). *Journal of the American Statistical Association* **85,** 640–644.

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression.* Chapman and Hall, London.

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

Hadi, A. S. (1992) Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society* **B54,** 761–771.

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.

Rousseeuw, P. J. (1984) Least median of squares regression. *Journal of the American Statistical Association* **79,** 871–880.

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

Rousseeuw, P. J. and van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* **85,** 633–639.

Weisberg, S. (1985) *Applied Linear Regression* (second edition). Wiley, New York.

Woodruff, D. L. and Rocke, D. M. (1992) Computation of minimum volume ellipsoid estimates using heuristic search. Technical report, Graduate School of Management, University of California at Davis.