J. R. Statist. Soc. B (1992)
54, *No.* 3, *pp.* 761–771

Identifying Multiple Outliers in Multivariate Data

By ALI S. HADI†

Cornell University, Ithaca, USA

[Received March 1990. Final revision May 1991]

SUMMARY

We propose a procedure for the detection of multiple outliers in multivariate data. Let X be an $n \times p$ data matrix representing n observations on p variates. We first order the n observations, using an appropriately chosen robust measure of outlyingness, then divide the data set into two initial subsets: a 'basic' subset which contains $p+1$ 'good' observations and a 'non-basic' subset which contains the remaining $n-p-1$ observations. Second, we compute the relative distance from each point in the data set to the centre of the basic subset, relative to the (possibly singular) covariance matrix of the basic subset. Third, we rearrange the n observations in ascending order accordingly, then divide the data set into two subsets: a basic subset which contains the first $p+2$ observations and a non-basic subset which contains the remaining $n-p-2$ observations. This process is repeated until an appropriately chosen stopping criterion is met. The final non-basic subset of observations is declared an outlying subset. The procedure proposed is illustrated and compared with existing methods by using several data sets. The procedure is simple, computationally inexpensive, suitable for automation, computable with widely available software packages, effective in dealing with masking and swamping problems and, most importantly, successful in identifying multivariate outliers.

Keywords: LEVERAGE POINTS; MAHALANOBIS DISTANCE; MASKING; MINIMUM VOLUME ELLIPSOID; ROBUST DISTANCE; SWAMPING

1. INTRODUCTION

This paper deals with the problem of identifying multiple outliers in multivariate data. Let X be an $n \times p$ matrix representing a random sample of size n from a p-dimensional population. Of interest is the detection of subset(s) of observations which are outlying in the p-dimensional scatter of points generated by X.

Classical outlier detection methods are powerful when the data contain only one outlier. However, the powers of these methods decrease drastically if more than one outlying observations are present in the data. This loss of power is usually due to what are known as the masking and swamping problems. In addition, these methods do not always succeed in detecting outliers, simply because they are affected by the observations that they are supposed to identify. Therefore, a method which avoids these problems is needed.

Let $D_i(C, V) = f(x_i - C, V)$, $i = 1, \ldots, n$, be an appropriate metric that measures the distance between the ith observation x_i and a location (centre) estimator C, relative to a measure of dispersion, V. Several forms of C and V are discussed in, for example, Rousseeuw and Leroy (1987), chapter 7, and in the references therein. The most commonly used form of $f(x_i - C, V)$ is given by

†*Address for correspondence*: Department of Economic and Social Statistics, 358 Ives Hall, Cornell University, Ithaca, NY 14853-3901, USA.

 0035-9246/92/54761 $2.00

$$D_i(C, V) = \sqrt{\{(x_i - C)^T V^{-1}(x_i - C)\}}, \qquad i = 1, \ldots, n. \tag{1.1}$$

The classical choices of C and V are respectively the arithmetic mean $\bar{X}$ and the sample covariance matrix S of the data set X, in which case equation (1.1) becomes

$$\mathrm{MD}_i = D_i(\bar{X}, S) = \sqrt{\{(x_i - \bar{X})^T S^{-1}(x_i - \bar{X})\}}, \qquad i = 1, \ldots, n. \tag{1.2}$$

This is known as the Mahalanobis distance. We refer to equation (1.2) as MD_i for simplicity.

A large value of MD_i may indicate that the corresponding observation is an outlier. However, two problems arise in practice. First, outliers do not necessarily have large values for MD_i. For example, a small cluster of outliers will attract $\bar{X}$ and will inflate S in its direction, yielding small values for MD_i. This problem is known as the masking problem because the presence of one outlier masks the appearance of another outlier.

Second, not all observations with large MD_i values are necessarily outliers. For example, a small cluster of outliers will attract $\bar{X}$ and inflate S in its direction and away from some other observations which belong to the pattern suggested by the majority of observations, thus yielding large MD_i values for these observations. This problem is known as the swamping problem.

The problems of masking and swamping are due to the fact that $\bar{X}$ and S are not robust. One way to avoid such problems is to use more robust estimators of the location and covariance matrix. Several such estimators have been suggested; see, for example, Campbell (1980), Stahel (1981), Donoho (1982), Hampel *et al.* (1986), Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990).

Rousseeuw (1985) uses the minimum volume ellipsoid (MVE) that covers at least half of the observations to construct robust estimators. The centre and covariance matrix of the observations included in the MVE are robust location and covariance matrix estimators. The advantage of MVE estimators is that they have a breakdown point of approximately 50% (Lopuhaä and Rousseeuw, 1991). However, it is computationally expensive, and it may not even be computationally feasible, to find the MVE. For an $n \times p$ data matrix X, if h is the integer part of $(n+1)/2$, then we need to compute the volumes of $n!/h!(n-h)!$ ellipsoids and to select the ellipsoid with the minimum volume. For example, for $n = 20$ there are 184756 such ellipsoids, and for $n = 30$ there are more than 155 million ellipsoids!

To deal with this computational difficulty, several algorithms have been suggested for approximating the MVE. One such algorithm is the resampling algorithm (Rousseeuw and Leroy, 1987). For other approximate algorithms, see, for example, Stahel (1981) and Donoho (1982). The resampling algorithm draws several subsamples each of size $p+1$; then for each subsample j we compute

$$D_i(C_j, S_j) = \sqrt{\{(x_i - C_j)^T S_j^{-1}(x_i - C_j)\}}, \qquad i = 1, \ldots, n, \tag{1.3}$$

where C_j and S_j denote the mean and covariance matrix for the jth subsample. Let m_j be the $100(h/n)$th percentile of the n values in equation (1.3). The volume of an ellipsoid based on C_j and S_j and containing h observations is proportional to $\{m_j^p \det(S_j)\}^{1/2}$. Let j be the subsample for which $m_j^p \det(S_j)$ is a minimum. The ellipsoid based on subsample j is used as an approximation to the MVE containing h observations. Rousseeuw and van Zomeren (1990) set h to the integer part of $(n+p+1)/2$ and suggest use of the robust distances

$$\mathrm{RD}_i = D_i(C_j, c_j S_j) = \sqrt{\{(x_i - C_j)^T (c_j S_j)^{-1}(x_i - C_j)\}}, \qquad i = 1, \ldots, n, \tag{1.4}$$

to identify outliers in the data set X. For simplicity we refer to equation (1.4) as RD_i. The constant $c_j = c_{np} m_j / \chi^2_{p,0.50}$ is a correction factor to obtain consistency when the data come from a multivariate normal distribution. On the basis of a simulation study for $p \leqslant 8$, Rousseeuw and van Zomeren (1990) use $c_{np} = \{1 + 15/(n-p)\}^2$.

Distance (1.4) is indeed a robust distance and, therefore, it is better than equation (1.2) in dealing with the problems of masking and swamping. However, three problems arise when applying equation (1.4) in practice. First, a decision has to be made on the number of subsamples. This decision can be made on the following probabilistic argument. Let m be the number of subsamples and let π be the probability that at least one of the m subsamples will contain no outliers. Since π depends on m and k, the number of outliers in the data, then it can be expressed as $\pi = f(m, k)$. Thus, m must be equal to or larger than $f^{-1}(\pi, k)$, where the inverse is with respect to m for fixed k.

The second, and more serious, problem that arises when applying equation (1.4) in practice is that equations (1.3) and (1.4) are based on the assumption that X is in the general position. (X is said to be in the general position when every subsample of size $p+1$ has rank p.) From the practical point of view, this is clearly unrealistic. If the rank of a subset j is less than p, then $\det(S_j) = 0$ and, hence, the volume of the corresponding ellipsoid is 0 and the distances RD_i in equation (1.4) cannot be computed. Rousseeuw and van Zomeren (1990) avoid this difficulty by simply omitting any subsample with (nearly) singular covariance matrix. A method which searches for a minimum volume ellipsoid and ignores ellipsoids with zero volumes seems to defeat its own purpose.

Third, even if all subsamples of size $p+1$ have rank p, it may happen that the covariance matrices for some subsamples have nearly zero determinants and hence the corresponding ellipsoids have nearly zero volumes. Subsequently, subsamples which have approximately the same nearly zero (minimum) volumes may have completely different ellipsoids, and hence equation (1.4) may give different results depending on which ellipsoid is used.

In this paper, we propose a procedure for approximating the MVE which is not subject to these problems. The procedure is easy to compute, does not depend on resampling, and hence gives a unique MVE, and works well even if S_j is singular.

2. PROPOSED PROCEDURE

We propose the following procedure for identifying multiple outliers in multivariate data.

2.1. *Step 0: Initial Ordering*

Initially rearrange the n observations in ascending order according to a suitably chosen robust distance. For example, we may use

$$D_i(C_R, S_R) = \sqrt{\{(x_i - C_R)^T S_R^{-1} (x_i - C_R)\}}, \qquad i = 1, \ldots, n, \tag{2.1}$$

where C_R and S_R are robust location and covariance matrix estimators. (Choices for C_R and S_R are given in Appendix A.) We then divide the observations into two initial subsets: one subset contains the first $p+1$ observations and the other subset contains

the last $n-p-1$ observations. We refer to these subsets as the 'basic' and 'non-basic' subset respectively.

2.2. *Step 1(a): Basic Subset of Full Rank*

If the basic subset is of full rank, compute

$$\sqrt{\{(x_i-C_b)^T S_b^{-1}(x_i-C_b)\}}, \qquad i=1, \ldots, n, \tag{2.2}$$

where C_b and S_b are the mean and covariance matrix of the basic subset.

2.3. *Step 1(b): Basic Subset Not of Full Rank*

If the basic subset is not of full rank, we first compute the eigenvalues of S_b, $\lambda_1 \geqslant \ldots \geqslant \lambda_p = 0$, and the matrix containing the corresponding set of normalized eigenvectors, V_b. Then we compute the distances

$$\sqrt{\{(x_i-C_b)^T V_b W_b V_b^T (x_i-C_b)\}}, \qquad i=1, \ldots, n, \tag{2.3}$$

where W_b is a diagonal matrix whose jth diagonal element is

$$w_j = \frac{1}{\max\{\lambda_j, \lambda_s\}}, \qquad j=1, \ldots, n, \tag{2.4}$$

and λ_s is the smallest non-zero eigenvalue of S_b.

2.4. *Step 2: Increase Size of Basic Subset*

Rearrange the observations in ascending order according to either expression (2.2) or expression (2.3) depending on whether S_b is or is not of full rank. Let r be the number of observations in the current basic subset. Divide the observations into two subsets: a basic subset containing the first $r+1$ observations and another subset containing the remaining $n-r-1$ observations.

2.5. *Step 3: Stopping Criterion*

Repeat steps 1 and 2 until a certain stopping criterion is met, then compute the robust distances

$$D_i(C_b, S_b) = \sqrt{\{(x_i-C_b)^T (c_b S_b)^{-1}(x_i-C_b)\}}, \qquad i=1, \ldots, n, \tag{2.5}$$

where $c_b = c_{npr} m_j / \chi^2_{p,0.50}$ is a correction factor to obtain consistency when the data come from a multivariate normal distribution. Our simulation results indicate that an appropriate small sample correction factor is $c_{npr} = \{1+r/(n-p)\}^2$, where r is the number of observations in the final basic subset. This factor increases with p for fixed n and decreases with n and r for fixed p, as it should.

2.6. *Rationale*

The logic behind this procedure is as follows. We initially rearrange observations in ascending order using distance (2.1) which is based on robust estimators of location

and scale (see Appendix A). In this way outliers are highly likely to appear at the end of the data and the initial basic subset is highly unlikely to contain outliers.

Step 1 uses this clean basic subset to compute robust estimators of the location and covariance matrix, successively. We then need to compute a distance measure from each observation to the centre of the observations contained in the basic subset. This distance is relative to the covariance matrix of the observations in the basic subset. If the basic subset is of full rank, then the corresponding covariance matrix is non-singular and the reason for using expression (2.2) is clear. It is similar to Mahalanobis distance (1.2) but uses more robust estimates of the location and covariance matrix. If the basic subset is not of full rank, we use the distance in expression (2.3), the rationale for which is given in Appendix B.

In step 2, we increase the size of the basic subset by one observation until a stopping criterion is met. In step 3 we have to decide on a stopping criterion. Two stopping criteria are

(a) stop when $\min\{D_i(C_b, S_b);\ i \in \text{non-basic subset}\} \geqslant c_\alpha$) or
(b) stop when the basic subset contains h observations, where h is to be defined shortly.

The critical value c_α can be chosen such that

$$\Pr[\min\{D_i(C_b, S_b);\ i \in \text{non-basic subset}\} \geqslant c_\alpha \mid X \text{ contains no outliers}] = 1 - \alpha.$$

However, c_α depends on the distribution of $D_i(C_b, S_b)$ which is clearly difficult to derive. Until we know how to compute c_α, we adopt rule (b).

Rule (b) requires $h-p$ iterations and distance (2.5) will be based on h observations. As h increases, distance (2.5) becomes more accurate but less robust. It becomes more accurate because, as h increases, the estimates of location and covariance matrix used in equation (2.5) will be based on more observations. It becomes less robust because, for fixed k, the probability that the basic subset will contain no outliers grows smaller as h grows larger. Obviously, there is a trade-off between accuracy and robustness. It is clear, however, that the upper bound on h is $n-k$. (In the unlikely event that the number of outliers, k, is known, we set $h=n-k$.) Since we do not know k in the examples of the next section, we set h to the integer part of $(n+p+1)/2$ so that our results are comparable with those of Rousseeuw and van Zomeren (1990).

Once the stopping criterion has been met, the location and covariance matrix estimator based on the observations included in the final basic set are used to compute equation (2.5). The observations with large values of distance (2.5) are then declared outliers. As mentioned earlier, since the distribution of $D_i(C_b, S_b)$ is difficult to derive, it is difficult to determine statistically how large is large. Therefore, we resort to graphical displays. In an index plot of $D_i(C_b, S_b)$, for example, outlying observations tend to appear far removed from the majority of the other observations in the data set.

3. EXAMPLES

In this section we investigate the effectiveness of our procedure empirically. To facilitate comparisons with existing procedures, we use the three data sets used by Rousseeuw and van Zomeren (1990).

3.1. *Example 1: Brain and Weight Data (Jerison, 1973)*

The brain and weight data set, which is taken from Rousseeuw and Leroy (1987), p. 57, contains two variables, log(brain weight) and log(body weight) for 28 species. It is part of a larger data set in Jerison (1973).

With $n=28$ and $p=2$, there are 3276 possible subsamples of size $p+1=3$ observations. The robust distance RD_i of equation (1.4) can be computed either by complete enumeration of the 3276 possible subsamples or by selecting subsamples of size 3. However, the number of all possible subsamples increases factorially with n and p, and for large values of n and p it is computationally expensive, and may not even be feasible, to search all possible subsamples.

In this data set, the covariance matrices of the four subsets {2, 9, 12}, {4, 12, 22}, {8, 19, 21} and {11, 17, 21} have determinants less than 10^{-8}. Thus, for each of these subsets S_j is essentially singular and hence its inverse essentially does not exist. Rousseeuw and van Zomeren (1990) avoid this difficulty by simply omitting any subsample with (nearly) singular covariance matrix.

The robust distance RD_i is computed by selecting 1500 subsamples of size 3. Of the 1500 selected subsamples the subset {1, 2, 22} has the minimum volume. The robust distance RD_i is based on this subset. Fig. 1 shows the index plots for the Mahalanobis distance MD_i and the robust distances RD_i and $D_i(C_\mathrm{b}, S_\mathrm{b})$. At the 5% level, the critical value is $(\chi^2_{2,0.975})^{1/2}=2.72$. Thus, the classical Mahalanobis distance declares only observation 25 (brachiosaurus) as an outlier, whereas both the robust distances RD_i and $D_i(C_\mathrm{b}, S_\mathrm{b})$ declare four observations as outliers (25, 6, 16 and 14, namely three dinosaurs with relatively small brains and the human with a relatively heavy brain). RD_i shows that observation 17 is on the boundary of the rejection region. (These cut-off points should be taken with a grain of salt because RD_i and $D_i(C_\mathrm{b}, S_\mathrm{b})$ follow a χ^2_p-distribution only approximately. The values in each column should be compared relative to each other. This comparison can be accomplished with graphical displays such as an index plot. Thus, for example, a visual inspection of the plots in Figs 1(b) and 1(c) indicates that only observations 25, 6 and 16 are outliers.)

3.2. *Example 2: Stack Loss Data (Brownlee, 1965)*

The stack loss data set consists of 21 observations on three explanatory and one response variables. This data set has been widely used to illustrate outliers and influential observations in linear regression. Here we use only the three explanatory variables. Of the 5985 possible subsamples of size 4 observations, there are 266 subsamples whose covariance matrices have determinants less than 10^{-8}.

A search of all the remaining 5719 subsets found two subsets ({7, 10, 14, 20} and {8, 10, 14, 20}) with the same minimum volume ($\det(S_j) = 472.9$, $m_j=3.87$, and the volume is proportional to $\{m_j^3 \det(S_j)\}^{1/2} = \sqrt{(3.87^3 \times 472.9)} = 165.56$). The reason that we find two subsets with the same volume here is that observations 7 and 8 are identical. However, this is not necessarily always the case. It is easy to construct data sets in which two different subsets have the same minimum volume, yet their corresponding ellipsoids have completely different shapes and even different orientations.

To compute RD_i by sampling, 2000 random subsamples were selected. (In the sampling process, 101 subsamples with singular covariance matrices were encountered and omitted.) Of the subsamples selected, the set {7, 10, 14, 20} has the

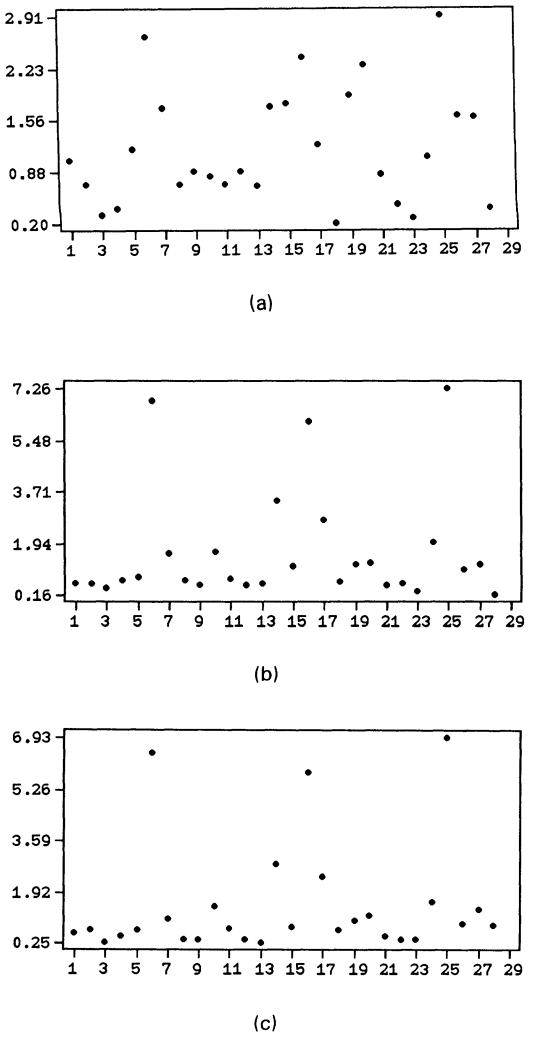

Fig. 1. Index plots for the brain and weight data: (a) index plot of MD_i; (b) index plot of RD_i; (c) index plot of $D_i(C_b, S_b)$

minimum volume. The index plots for MD_i, RD_i and $D_i(C_b, S_b)$ are given in Fig. 2.

At the 5% level, the critical value is $(\chi^2_{3,0.975})^{1/2} = 3.06$. Thus, the Mahalanobis distance fails to detect any of the many outliers known to be present in this data set (the largest value is $MD_{17} = 2.70$). Both RD_i and $D_i(C_b, S_b)$ declare four observations as outliers (2, 1, 3 and 21, in this order).

3.3. *Example 3: Hawkins, Bradu and Kass Data*

The data from Table 4 of Hawkins *et al.* (1984) is a constructed data set with $n = 75$ and $p = 3$. It provides a good example of the masking effect. The first three variables are constructed so that the first 14 observations are outliers. There are 1215450 distinct subsets of size 4. Because of this large number of subsets, we compute RD_i by

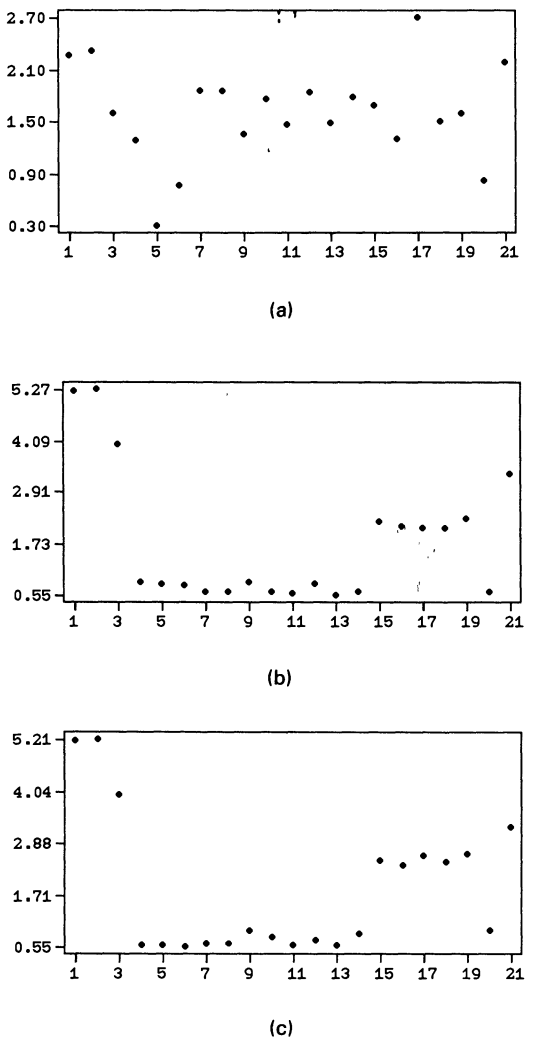

Fig. 2. Index plots for the stack loss data: (a) index plot of MD_i; (b) index plot of RD_i; (c) index plot of $D_i(C_b, S_b)$

selecting 2000 random subsets. The index plots for MD_i, RD_i and $D_i(C_b, S_b)$ are given in Fig. 3. At the 5% level, the critical value is $(\chi^2_{3,0.975})^{1/2} = 3.06$. The Mahalanobis distance reveals only two of the 14 outliers (observations 12 and 14). These two outliers mask all the other 12 outliers. All the 14 outliers are unmasked by both RD_i and $D_i(C_b, S_b)$.

4. SUMMARIZING REMARKS AND CONCLUSIONS

In this paper we have proposed a procedure for identifying multiple outliers in multivariate data. The classical Mahalanobis distance is clearly not effective in identifying multiple outliers as it suffers from masking and swamping problems. We

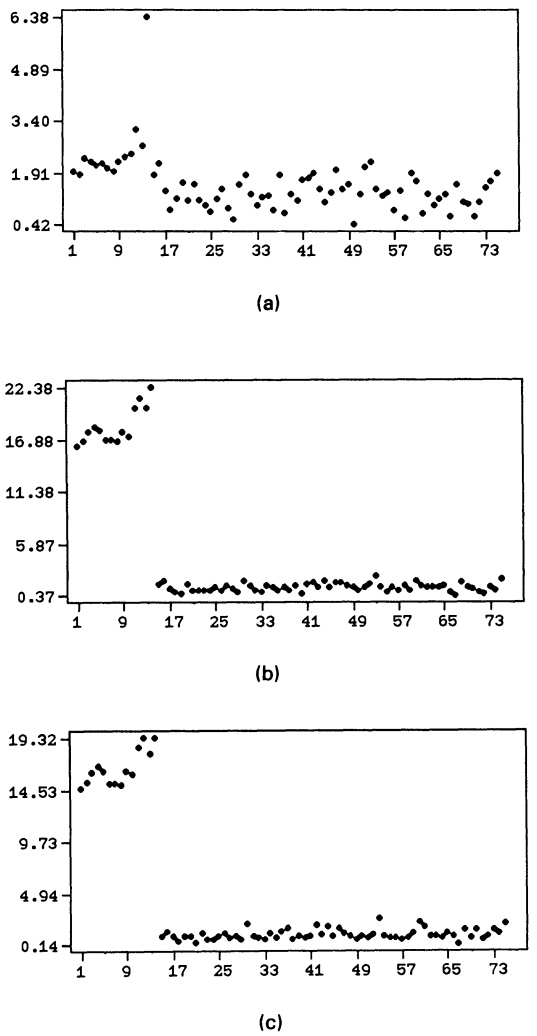

Fig. 3. Index plots for the Hawkins, Bradu and Kass data: (a) index plot of MD_i; (b) index plot of RD_i; (c) index plot of $D_i(C_b, S_b)$

have seen in the examples of Section 3 that RD_i and $D_i(C_b, S_b)$ are equally effective in identifying multivariate outliers and in dealing with masking and swamping problems. This is because they are based on robust estimators of location and covariance matrix. The actual breakdown point of S_b depends on the initial estimators. Because the h observations on which S_b is based are highly unlikely to contain outliers, S_b is more robust than S. The breakdown points of C_b and S_b are open problems. $D_i(C_b, S_b)$, however, has the following advantages over RD_i.

(a) RD_i requires substantial computing time and, because it depends on resampling, it is not as suitable for automation. Our procedure is simple, computationally inexpensive and can be computed with widely available software packages.

(b) The mean vector C_j and the covariance matrix S_j are based on only $p+1$ observations, whereas C_b and S_b are based on $h = (n+p+1)/2$ observations. Hence, C_b and S_b are more accurate than C_j and S_j.

(c) Also, because the covariance matrix S_j is based on $p+1$ observations, it is highly likely that one of the subsamples will yield a matrix S_j with a very small determinant. As a result, S_b is computationally more stable than S_j.

(d) In practice, subsamples with rank deficient covariance matrices are quite common; for example, in the stack loss data there are 266 subsets of size 4 with S_j singular. If one of these subsamples is selected, the volume of the corresponding ellipsoid is 0 and RD_i is not computable. In cases where a basic subset is rank deficient, we have suggested in expression (2.3) a quantity which measures the distance between C_b and each point in the data set relative to the singular covariance matrix S_b.

ACKNOWLEDGEMENTS

I am grateful to Professor Peter J. Rousseeuw and Professor Steven J. Schwager for their helpful comments on an earlier version of this paper. The present form of the paper greatly benefited from comments and suggestions made by the Editor, an associate editor and two referees.

APPENDIX A: ROBUST LOCATION AND SCALE ESTIMATORS

In step 0, we order observations by using distance (2.1) which is based on robust location and scale estimators C_R and S_R. These estimators are obtained by first computing $D_i(C_M, S_M)$, where C_M is a vector containing the co-ordinatewise medians and

$$S_M = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - C_M)(x_i - C_M)^T.$$

Now rearrange the observations in ascending order according to $D_i(C_M, S_M)$, define the weight function

$$v_i = \begin{cases} 1, & \text{if } i \leqslant \text{integer part of } (n+p+1)/2, \\ 0, & \text{otherwise,} \end{cases}$$

and compute equation (2.1) by setting $C_R = C_v$ and $S_R = S_v$, where C_v and S_v are defined by

$$C_v = \frac{\sum_{i=1}^{n} v_i x_i}{\sum_{i=1}^{n} v_i} \quad \text{and} \quad S_v = \frac{\sum_{i=1}^{n} v_i (x_i - C_v)(x_i - C_v)^T}{\sum_{i=1}^{n} v_i - 1}.$$

Thus in step 0 we rearrange observations in ascending order according to $D_i(C_v, S_v)$.

APPENDIX B: RATIONALE FOR DISTANCE (2.3)

If the basic subset is not of full rank, S_b is not invertible and expression (2.2) cannot be computed. A distance between each observation in the data set and the centre of the basic subset, C_b, relative to S_b exists regardless of whether S_b is or is not invertible, but we simply

cannot measure it by using expression (2.2). A measure of such a distance is needed. We measure this distance by expression (2.3). We note that S_b can be written as $S_b = V_b \Lambda_b V_b^T$, where Λ_b is a diagonal matrix containing the eigenvalues of S_b. If S_b is non-singular, expression (2.2) can be written as

$$\sqrt{\{(x_i - C_b)^T V_b \Lambda_b^{-1} V_b^T (x_i - C_b)\}}, \qquad i = 1, \ldots, n, \tag{B.1}$$

and it can be thought of as the square root of a weighted sum of squares of the elements of $V_b^T(x_i - C_b)$. The weights are the diagonal elements of Λ_b^{-1}. The problem is that when S_b is singular $\lambda_p = 0$ and $1/\lambda_p$ does not exist. One way out of this problem is to replace the weight $1/\lambda_j$ by w_j in equation (2.4). In this case expression (B.1) becomes expression (2.3). Note that the weight w_j is inversely related to λ_j, as it should be. Also, when $j = 1$, $w_1 = 1/\lambda_1$, as it should be.

When S_b is non-singular, expressions (2.2), (2.3) and (B.1) are equivalent but expression (2.2) is preferable because it does not require computations of the eigenvalues and eigenvectors of S_b.

Another way of measuring the distance between each observation in the data set and the centre of the basic subset, C_b, relative to S_b, is to project the observations in the data set on to the direction where the observations in the basic subset is the least variable. If S_b is singular, then at least one eigenvalue of S_b is 0. Let V_2 be the submatrix of V_b which contains the normalized eigenvectors corresponding to the zero eigenvalue(s) of S_b. Then V_2 is the direction of least variability. Because $V_2^T V_2 = I$, the corresponding projection matrix is $V_2(V_2^T V_2)^{-1} V_2^T = V_2 V_2^T$. Therefore, $V_2 V_2^T (x_i - C_b)$ is a projection of $(x_i - C_b)$ on V_2. The norm of this projection is

$$\sqrt{\{(x_i - C_b)^T V_2 V_2^T V_2 V_2^T (x_i - C_b)\}} = \sqrt{\{(x_i - C_b)^T V_2 V_2^T (x_i - C_b)\}}. \tag{B.2}$$

Thus, expression (B.2) is simply the norm of the projection of each observation on the eigenvector corresponding to the zero eigenvalue of V_2. Expression (B.2) is inferior to expression (2.3) because it ignores other co-ordinates, but it is computationally better because it requires the computation of only $p - \text{rank}(S_b)$ eigenvectors without the need to compute the eigenvalues.

REFERENCES

Brownlee, K. A. (1965) *Statistical Theory and Methodology in Science and Engineering*, 2nd edn. New York: Wiley.

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

Donoho, D. L. (1982) Breakdown properties of multivariate location estimators. *PhD Dissertation*. Harvard University, Boston.

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

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

Jerison, H. J. (1973) *Evolution of the Brain and Intelligence*. New York: Academic Press.

Lopuhaä, H. P. and Rousseeuw, P. J. (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, **19**, 229–248.

Rousseeuw, P. J. (1985) Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications* (eds W. Grossmann, G. Pflug, I. Vincze and W. Wertz), vol. B, pp. 283–297. Dordrecht: Reidel.

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

Rousseeuw, P. J. and van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points (with comments). *J. Am. Statist. Ass.*, **85**, 633–651.

Stahel, W. A. (1981) Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. *PhD Thesis*. Eidgenossiches Technische Hochschule, Zurich.