Unmasking Multivariate Outliers and Leverage Points

PETER J. ROUSSEEUW and BERT C. VAN ZOMEREN*

Detecting outliers in a multivariate point cloud is not trivial, especially when there are *several* outliers. The classical identification method does not always find them, because it is based on the sample mean and covariance matrix, which are themselves affected by the outliers. That is how the outliers get *masked*. To avoid the masking effect, we propose to compute distances based on very robust estimates of location and covariance. These robust distances are better suited to expose the outliers.

In the case of regression data, the classical least squares approach masks outliers in a similar way. Also here, the outliers may be unmasked by using a highly robust regression method. Finally, a new display is proposed in which the robust regression residuals are plotted versus the robust distances. This plot classifies the data into regular observations, vertical outliers, good leverage points, and bad leverage points. Several examples are discussed.

KEY WORDS: Breakdown point; Leverage diagnostic; Mahalanobis distance; Minimum volume ellipsoid; Residual plot.

1. IDENTIFICATION OF MULTIVARIATE OUTLIERS

Outliers are observations that do not follow the pattern of the majority of the data. Outliers in a multivariate point cloud can be hard to detect, especially when the dimension p exceeds 2, because then we can no longer rely on visual perception. A classical method is to compute the Mahalanobis distance

$$\mathrm{MD}_i = \sqrt{(\mathbf{x}_i - T(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i - T(\mathbf{X}))^t} \qquad (1)$$

for each point $\mathbf{x}_i$. Here, $T(\mathbf{X})$ is the arithmetic mean of the data set $\mathbf{X}$ and $\mathbf{C}(\mathbf{X})$ is the usual sample covariance matrix. The distance MD_i should tell us how far $\mathbf{x}_i$ is from the center of the cloud, taking into account the shape of the cloud as well. It is well known that this approach suffers from the *masking effect,* by which multiple outliers do not necessarily have a large MD_i. This is due to the fact that $T(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$ are not robust: a small cluster of outliers will attract $T(\mathbf{X})$ and will inflate $\mathbf{C}(\mathbf{X})$ in its direction. Therefore, it seems natural to replace $T(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$ in (1) by robust estimators.

Campbell (1980) proposed to insert M estimators for $T(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$, which marked an important improvement. Unfortunately, the breakdown point of M estimators (i.e., the fraction of outliers they can tolerate) is at most $1/(p + 1)$, so it goes down when there are more coordinates in which outliers can occur [see, e.g., chapter 5 of Hampel, Ronchetti, Rousseeuw, and Stahel (1986)].

As a further step, one may consider estimators of multivariate location and covariance that have a high breakdown point. The first such estimator was proposed by Stahel (1981) and Donoho (1982). Here we will use the minimum volume ellipsoid estimator (MVE) introduced by Rousseeuw (1985). For $T(\mathbf{X})$ we take the center of the minimum volume ellipsoid covering half of the observations, and $\mathbf{C}(\mathbf{X})$ is determined by the same ellipsoid (multiplied by a correction factor to obtain consistency at multinormal distributions). A technical definition of the MVE estimator is given in the Appendix, together with two approximate algorithms for its computation. We denote by RD_i the robust distances obtained by inserting the MVE estimates for $T(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$ in (1).

Figure 1 illustrates the distinction between classical and robust estimates. It is a log–log plot of brain weight versus body weight for 28 species. The raw data (before taking logarithms) can be found in Rousseeuw and Leroy (1987, p. 58), where they were used for a different purpose. In Figure 1 we see that the majority of the data follow a clear pattern, with a few exceptions. In the lower right region there are three dinosaurs (observations 6, 16, and 25) with a small brain and a heavy body, and in the upper left area we find the human and the rhesus monkey (observations 14 and 17) with a relatively high brain weight. The 97.5% tolerance ellipse obtained from the classical estimates (dashed line) is blown up by these outliers, and contains all animals but the largest dinosaur. The tolerance ellipse based on the MVE is much narrower (solid line) and does not include the outliers.

The second column of Table 1 shows the classical Mahalanobis distances for these observations. The only outlier outside the tolerance ellipse (number 25) yields the only MD_i exceeding the cutoff value $\sqrt{\chi^2_{2,.975}} = 2.72$. On the other hand, the robust distances RD_i in the rightmost column do identify the exceptional observations (all values larger than 2.72 are underscored).

Of course, in two dimensions we can still look at a plot of the data to find the outliers. Algorithms really become necessary in three and more dimensions. For instance, consider the explanatory variables of the stackloss data (Brownlee 1965). This point cloud (with $p = 3$ and $n = 21$) contains several outliers. The second column of Table 2 gives the classical MD_i, all of which stay beneath $\sqrt{\chi^2_{3,.975}} = 3.06$. The largest MD_i (of observation 17) is only 2.70. The robust distances in the next column, however, clearly pinpoint four outliers (cases 1, 2, 3, and 21).

* Peter J. Rousseeuw is Professor, U.I.A., Vesaliuslaan 24, B-2650 Edegem, Belgium. Bert C. van Zomeren is Teaching Assistant, Faculty of Mathematics and Informatics, Delft University of Technology, Julianalaan 132, 2628 BL Delft, The Netherlands. The authors are grateful to David Donoho, John Tukey, and two referees for interesting discussions and helpful comments.

Journal of the American Statistical Association
September 1990, Vol. 85, No. 411, Theory and Methods

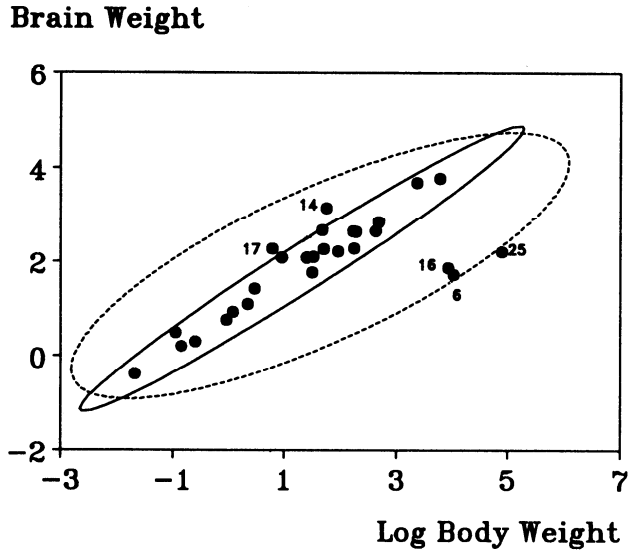

Figure 1. Plot of Log Brain Weight Versus Log Body Weight, With 97.5% Tolerance Ellipse Based on the Classical Mean and Covariance (dashed line) and on the Robust Estimator (solid line).

The data set of Hawkins, Bradu, and Kass (1984, table 4) yields a prime example of the masking effect. The first three variables form a point cloud with $n = 75$ and $p = 3$. It is known that cases 1 to 14 are outliers, but the classical MD_i in Table 3 do not reveal this. The only MD_i larger than $\sqrt{\chi^2_{3,.975}} = 3.06$ belong to observations 12 and 14, which mask all the others. On the other hand, the robust distances in the same table do expose the 14 outliers in a single blow.

Table 1. Classical Mahalanobis Distances (MD_i) and Robust Distances (RD_i) for the Brain Weight Data

i	MD_i	RD_i
1	1.01	.54
2	.70	.54
3	.30	.40
4	.38	.63
5	1.15	.74
6	2.64	6.83
7	1.71	1.59
8	.71	.64
9	.86	.48
10	.80	1.67
11	.69	.69
12	.87	.50
13	.68	.52
14	1.72	3.39
15	1.76	1.14
16	2.37	6.11
17	1.22	2.72
18	.20	.67
19	1.86	1.19
20	2.27	1.24
21	.83	.47
22	.42	.54
23	.26	.29
24	1.05	1.95
25	2.91	7.26
26	1.59	1.04
27	1.58	1.19
28	.40	.75

NOTE: Distances exceeding the cutoff value $\sqrt{\chi^2_{2,.975}} \approx 2.72$ are underscored.

Table 2. Mahalanobis Distances (MD_i) and Robust Distances (RD_i) for the Stackloss Data, Along With the Diagonal Elements of the Hat Matrix

i	MD_i	RD_i	h_{ii}
1	2.25	5.23	.30
2	2.32	5.27	.32
3	1.59	4.01	.17
4	1.27	.84	.13
5	.30	.80	.05
6	.77	.78	.08
7	1.85	.64	.22
8	1.85	.64	.22
9	1.36	.83	.14
10	1.75	.64	.20
11	1.47	.58	.16
12	1.84	.79	.22
13	1.48	.55	.16
14	1.78	.64	.21
15	1.69	2.23	.19
16	1.29	2.11	.13
17	2.70	2.07	.41
18	1.50	2.09	.16
19	1.59	2.29	.17
20	0.81	.64	.08
21	2.18	3.30	.28

NOTE: Distances exceeding the cutoff value $\sqrt{\chi^2_{3,.975}} \approx 3.06$ are underscored.

The robust estimates and distances have been computed by means of a Fortran 77 program, which can be obtained from us. The computation time is of course larger than that of the classical method, but it is quite feasible (even on a PC) and the user obtains much information at once. We would like to stress that the user does not have to choose any tuning constants in advance and, in fact, the examples in this article were obtained from routine application of the program. Note that we do not necessarily want to *delete* the outliers; it is only our purpose to *find* them, after which the user may decide whether they are to be kept, deleted, or corrected, depending on the situation.

Remark. Detecting outliers turns out to be hardest when n/p is relatively small. In such a case a few data points may be nearly collinear by chance, thereby completely determining the MVE. This is caused by the emptiness of multivariate space (the "curse of dimensionality"). As a rule of thumb we recommend applying the MVE when there are at least five observations per dimension, so $n/p > 5$.

Robust covariance matrices can be used to detect outliers in several kinds of multivariate analysis, such as principal components (Campbell 1980; Devlin, Gnanadesikan, and Kettenring 1981) and canonical correlation and correspondence analysis (Karnel 1988).

2. IDENTIFICATION OF LEVERAGE POINTS IN REGRESSION

In linear regression the cases are of the type $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ is p-dimensional and the response y_i is one-dimensional. Cases for which $\mathbf{x}_i$ is far away from the bulk of the $\mathbf{x}_i$ in the data we call *leverage points*. Leverage points occur frequently when the $\mathbf{x}_i$ are observational, unlike "designed" situations with fixed $\mathbf{x}_i$. Leverage points may be

Table 3. Mahalanobis Distances (MD_i) and Robust Distances (RD_i) for the Hawkins–Bradu–Kass Data, Along With the Diagonal Elements of the Hat Matrix

i	MD_i	RD_i	h_{ii}	i	MD_i	RD_i	h_{ii}
1	1.92	16.20	.063	39	1.27	1.34	.035
2	1.86	16.62	.060	40	1.11	.55	.030
3	2.31	17.65	.086	41	1.70	1.48	.052
4	2.23	18.18	.081	42	1.77	1.74	.055
5	2.10	17.82	.073	43	1.87	1.18	.061
6	2.15	16.80	.076	44	1.42	1.82	.041
7	2.01	16.82	.068	45	1.08	1.25	.029
8	1.92	16.44	.063	46	1.34	1.70	.038
9	2.22	17.71	.080	47	1.97	1.65	.066
10	2.33	17.21	.087	48	1.42	1.37	.041
11	2.45	20.23	.094	49	1.57	1.27	.047
12	3.11	21.14	.144	50	.42	.83	.016
13	2.66	20.16	.109	51	1.30	1.19	.036
14	6.38	22.38	.564	52	2.08	1.61	.072
15	1.82	1.54	.058	53	2.21	2.41	.079
16	2.15	1.88	.076	54	1.41	1.26	.040
17	1.39	1.03	.039	55	1.23	.66	.034
18	.85	.73	.023	56	1.33	1.21	.037
19	1.15	.59	.031	57	.83	.93	.023
20	1.59	1.49	.048	58	1.40	1.31	.040
21	1.09	.87	.030	59	.59	.96	.018
22	1.55	.90	.046	60	1.89	1.89	.062
23	1.09	.94	.029	61	1.68	1.31	.051
24	.97	.83	.026	62	.76	1.22	.021
25	.80	1.26	.022	63	1.29	1.17	.036
26	1.17	.86	.032	64	.97	1.14	.026
27	1.45	1.35	.042	65	1.15	1.40	.031
28	.87	1.00	.024	66	1.30	.78	.036
29	.58	.72	.018	67	.63	.37	.019
30	1.57	1.97	.047	68	1.55	1.64	.046
31	1.84	1.43	.059	69	1.07	1.17	.029
32	1.31	.95	.036	70	1.00	1.04	.027
33	.98	.73	.026	71	.64	.64	.019
34	1.18	1.42	.032	72	1.05	.52	.028
35	1.24	1.26	.034	73	1.47	1.14	.043
36	.85	.86	.023	74	1.65	.96	.050
37	1.83	1.26	.059	75	1.90	1.99	.062
38	.75	.92	.021				

NOTE: Distances exceeding the cutoff value $\sqrt{\chi^2_{3,.975}} \approx 3.06$ are underscored.

quite difficult to detect, however, when the $\mathbf{x}_i$ have dimension higher than 2, because then we are exactly in the situation described previously in Section 1.

In the usual multiple linear regression model given by $\mathbf{y} = \mathbf{X}\theta + \mathbf{e}$ people often use the diagonal elements of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ as diagnostics to identify leverage points. Unfortunately, the hat matrix, like the classical Mahalanobis distance, suffers from the masking effect. This can be explained by realizing that there exists a monotone relation between the h_{ii} and the MD_i of the $\mathbf{x}_i$:

$$h_{ii} = \frac{(MD_i)^2}{n-1} + \frac{1}{n}. \tag{2}$$

Therefore, the h_{ii} do not necessarily detect the leverage points, contrary to what is commonly believed. Many authors even *define* leverage in terms of h_{ii} which, in our opinion, confuses cause and effect: the cause is the fact that some $\mathbf{x}_i$ are outlying, whereas the h_{ii} are merely some (unreliable) diagnostics trying to find those points. As an illustration let us look at Table 2, which shows the h_{ii} for the stackloss data. The largest h_{ii} belongs to observation 17, whereas the RD_i identify observations 1, 2, 3, and 21. Another example is the Hawkins–Bradu–Kass data set (Table 3). We know that the first 14 observations are leverage points, but only 12, 13, and 14 have large h_{ii}. Therefore, we propose to use the robust distances of the $\mathbf{x}_i$ as *leverage diagnostics*, because they are less easily masked than the h_{ii}.

Saying that $(\mathbf{x}_i, y_i)$ is a leverage point refers only to the outlyingness of $\mathbf{x}_i$ but does not take the response y_i into account. If $(\mathbf{x}_i, y_i)$ lies far from the plane corresponding to the majority of the data, we say that it is a *bad* leverage point. Such a point is very harmful because it attracts or even tilts the classical least squares regression (hence the word "leverage"). On the other hand, if $(\mathbf{x}_i, y_i)$ does fit the linear relation it will be called a *good* leverage point, because it improves the precision of the regression coefficients.

To distinguish between good and bad leverage points we have to consider y_i as well as $\mathbf{x}_i$, and we also need to know the linear pattern set by the majority of the data. This calls for a high-breakdown regression estimator, such as least median of squares (LMS), defined by

$$\underset{\theta}{\text{minimize}} \ \underset{i=1,\ldots,n}{\text{median}} \ r_i^2(\hat{\theta}) \tag{3}$$

(Rousseeuw 1984), where $r_i(\hat{\theta}) = y_i - \mathbf{x}_i\hat{\theta}$ is the residual of the ith observation. The LMS estimate $\hat{\theta}$ is affine equi-

variant and has the maximal breakdown point. After computing $\hat{\theta}$ one also calculates the corresponding scale estimate, given by

$$\hat{\sigma} = k\sqrt{\underset{i=1,\ldots,n}{\text{median}}\, r_i^2(\hat{\theta})} \tag{4}$$

where k is a positive constant. The standardized LMS residuals $r_i/\hat{\sigma}$ can then be used to indicate *regression outliers*, that is, points that deviate from the linear pattern of the majority (Rousseeuw 1984).

Figure 2 illustrates our terminology in an example of simple regression. The majority of the data are regular observations, indicated by (a). Points (b) and (d) deviate from the linear pattern and hence are called regression outliers, but (c) is not. Both (c) and (d) are leverage points, because their $\mathbf{x}_i$ value is outlying. Therefore, we say that (c) is a good leverage point and (d) is a bad leverage point. The observation (b) is called a *vertical outlier*, because it is a regression outlier but not a leverage point.

The robust distances in Tables 2 and 3 indicate leverage points but cannot distinguish between good and bad ones, because the y_i are not used. On the other hand, the LMS residual plots in chapter 3 of Rousseeuw and Leroy (1987) pinpoint regression outliers without telling which ones are leverage points. Therefore, it seems like a good idea to construct a new display in which the robust residuals $r_i/\hat{\sigma}$ are plotted versus the robust distances RD_i. In Figure 3 this is done for the stackloss data. Points to the right of the vertical borderline through $\sqrt{\chi^2_{3,.975}} = 3.06$ are leverage points, whereas points outside the horizontal tolerance band $[-2.5, 2.5]$ are regression outliers. In this example the four points with the largest RD_i are also regression outliers, so they are bad leverage points. Figure 3 also contains a vertical outlier (observation 4), which is a regression outlier with $\text{RD}_i < \sqrt{\chi^2_{3,.975}}$. Our cutoff values are to some extent arbitrary, but in the plot we can recognize the boundary cases: observation 21 is not very far away in x-space, whereas case 2 is only a mild regression outlier.

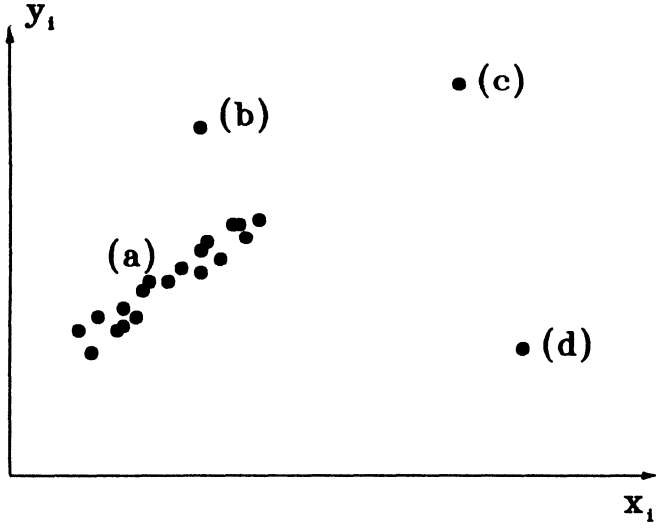

Figure 2. Simple Regression Example With (a) Regular Observations, (b) Vertical Outlier, (c) Good Leverage Point, and (d) Bad Leverage Point.

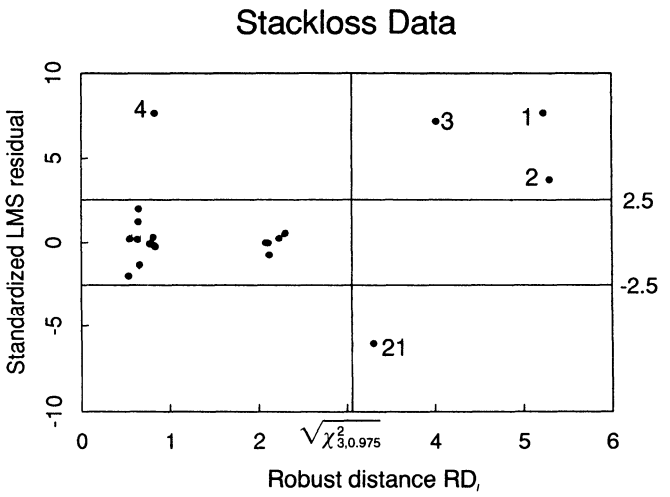

Figure 3. Plot of Robust Residuals Versus Robust Distances RD_i for the Stackloss Data.

A referee has asked to compare this display with its classical counterpart, which would plot the usual least squares residuals versus the nonrobust Mahalanobis distances MD_i. This plot is given in Figure 4 for the same data. It does not reveal any leverage points or regression outliers, because all of the points stay between the lines and only observations 21 and 17 come close to being identified. Because of (2), things would not improve when replacing MD_i by h_{ii}.

Figure 5 is the plot of robust residuals versus robust distances for the Hawkins–Bradu–Kass data. It immediately shows that there are 14 leverage points, of which 4 are good and 10 are bad. A glance at Figure 5 reveals the important features of these data, which are hard to discover otherwise. This type of plot presents a visual classification of the data into four categories: the regular observations with small RD_i and small $r_i/\hat{\sigma}$, the vertical outliers with small RD_i and large $r_i/\hat{\sigma}$, the good leverage points with large RD_i and small $r_i/\hat{\sigma}$, and the bad leverage points with large RD_i and large $r_i/\hat{\sigma}$. Note that a single

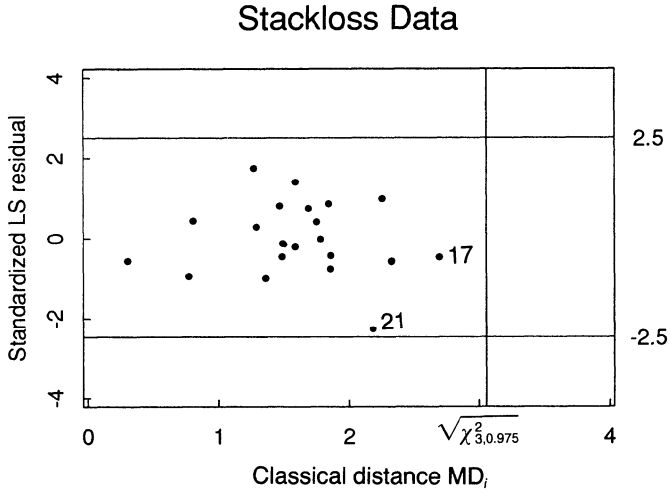

Figure 4. Plot of Least Squares Residuals Versus Classical Mahalanobis Distances MD_i for the Stackloss Data.

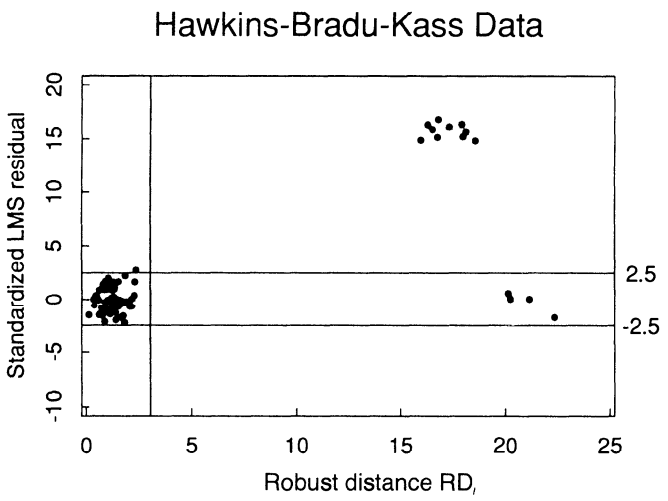

Figure 5. Plot of Robust Residuals Versus Robust Distances RD$_i$ for the Hawkins–Bradu–Kass Data.

diagnostic can never be sufficient for this fourfold classification!

Robust residuals may be used to assign weights to observations or to suggest data transformations (Carroll and Ruppert 1988; Rousseeuw and Leroy 1987). They are much better suited to this than least squares residuals, because least squares tries to produce normal-looking residuals even when the data themselves are not normal. The combination of the robust residuals with the RD$_i$ also offers another advantage. As pointed out by Atkinson (1986), it may sometimes happen that the LMS regression produces a relatively large residual at a *good* leverage point, because of small variations in the regression coefficients. The amplitude of this effect is roughly proportional to the RD$_i$, so the problem can only occur in the section on the right side of our new display. This is a distinct improvement over the usual plot of standardized residuals versus the index of the observation, where one does not see whether a given residual corresponds to an $\mathbf{x}_i$ at the center or to a leverage point.

3. CONCLUSIONS AND OUTLOOK

In this article we have proposed using distances based on high-breakdown estimators to detect outliers in a multivariate point cloud. This is in line with our previous suggestion to identify regression outliers by looking at residuals from a high-breakdown fit. Combining these tools leads to the robust diagnostic plot of Figures 3 and 5.

Although we do not claim this approach to be a panacea, it has worked very well for detecting outliers in many real-data examples not described here. Our general impression is that most data sets are further away from the usual assumptions (multivariate normality, approximate linearity) than is commonly assumed. In actual practice our methods have yielded some new and surprising results, for example, in a consulting firm fitting economic models to stock exchange data. Another application was to mining (Chork, in press), in which the outliers reflect mineralizations hidden below the surface, so their detection is the most important part of the analysis.

We would like to stress that we are *not* advocating that one simply remove the outliers. Instead we consider our plots of robust residuals and/or distances as a mere starting point of the analysis. In some cases the plots may tell us to change the model. In other cases we may be able to go back to the original data and explain where the outliers come from and, perhaps, to correct their values.

For the moment we are still carrying out simulations to compare different algorithms, study the distribution of robust distances, and so on. It turns out that it does not matter so much *which* high-breakdown estimator is used when the purpose is to detect outliers, because then statistical efficiency is less important than robustness.

Further research is needed to address situations where some of the explanatory variables are discrete, such as 0–1 dummies. The same is true for functionally related explanatory variables (e.g., polynomial terms), because then one cannot expect the majority of the $\mathbf{x}_i$ to form a roughly ellipsoidal shape. Nonlinear regression with high breakdown point has been addressed by Stromberg (1989).

Presently we are developing a program called ROMA (which stands for RObust Multivariate Analysis), incorporating both robust regression and robust location/covariance, as well as other techniques such as robust principal components.

Finally, we would like to apologize to all of the people whose work we did not cite. We did not attempt to write a review article (nor was it originally meant to be a discussion paper). Some reviews of the relevant literature on outliers and robustness can be found in Beckman and Cook (1983), Gnanadesikan (1977), Hampel et al. (1986), and Rousseeuw and Leroy (1987).

APPENDIX: METHODS AND ALGORITHMS

Suppose that we have a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of n points in p dimensions and we want to estimate its "center" and "scatter" by means of a row vector $T(\mathbf{X})$ and a matrix $\mathbf{C}(\mathbf{X})$. We say that the estimators T and $\mathbf{C}$ are *affine equivariant* when

$$T(\mathbf{x}_1\mathbf{A} + \mathbf{b}, \ldots, \mathbf{x}_n\mathbf{A} + \mathbf{b}) = T(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{A} + \mathbf{b}$$

and

$$\mathbf{C}(\mathbf{x}_1\mathbf{A} + \mathbf{b}, \ldots, \mathbf{x}_n\mathbf{A} + \mathbf{b}) = \mathbf{A}'\mathbf{C}(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{A} \quad \text{(A.1)}$$

for any row vector $\mathbf{b}$ and any nonsingular p-by-p matrix $\mathbf{A}$. The sample mean and the sample covariance matrix

$$T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \quad \text{and}$$

$$\mathbf{C}(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - T(\mathbf{X}))'(\mathbf{x}_i - T(\mathbf{X})) \quad \text{(A.2)}$$

are affine equivariant but not robust, because even a single outlier can change them to an arbitrary extent.

The minimum volume ellipsoid estimator (MVE) is defined as the pair $(T, \mathbf{C})$, where $T(\mathbf{X})$ is a p-vector and $\mathbf{C}(\mathbf{X})$ is a positive-semidefinite p-by-p matrix such that the determinant of $\mathbf{C}$ is minimized subject to

$$\#\{i;\ (\mathbf{x}_i - T)\mathbf{C}^{-1}(\mathbf{x}_i - T)' \le a^2\} \ge h \quad \text{(A.3)}$$

where $h = [(n + p + 1)/2]$ in which $[q]$ is the integer part of q. The number a^2 is a fixed constant, which can be chosen as $\chi^2_{p,.50}$ when we expect the majority of the data to come from a

normal distribution. For small samples one also needs a factor $c^2_{n,p}$, which depends on n and p. The MVE has a breakdown point of nearly 50%, which means that $T(\mathbf{X})$ will remain bounded and the eigenvalues of $\mathbf{C}(\mathbf{X})$ will stay away from zero and infinity when less than half of the data are replaced by arbitrary values (see, e.g., Lopuhaä and Rousseeuw, in press). The robust distances are defined relative to the MVE:

$$\mathrm{RD}_i = \sqrt{(\mathbf{x}_i - T(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i - T(\mathbf{X}))^t}. \quad (A.4)$$

One can then compute a weighted mean,

$$T_1(\mathbf{X}) = \left(\sum_{i=1}^n w_i\right)^{-1} \sum_{i=1}^n w_i \mathbf{x}_i, \quad (A.5)$$

and a weighted covariance matrix,

$$\mathbf{C}_1(\mathbf{X}) = \left(\sum_{i=1}^n w_i - 1\right)^{-1} \sum_{i=1}^n (\mathbf{x}_i - T_1(\mathbf{X}))^t(\mathbf{x}_i - T_1(\mathbf{X})) \quad (A.6)$$

where the weights $w_i = w(\mathrm{RD}_i)$ depend on the robust distances. It can be shown that T_1 and $\mathbf{C}_1$ have the same breakdown point as the initial T and $\mathbf{C}$ when the weight function w vanishes for large RD_i [see sec. 5 of Lopuhaä and Rousseeuw (in press)].

The MVE method can still be used when $p = 1$, in which case it yields the midpoint and the length of the shortest half. The midpoint converges merely as $n^{-1/3}$ (Rousseeuw 1984), whereas the length converges as $n^{-1/2}$ (Grübel 1988). The influence function and finite-sample behavior of the latter were studied by Rousseeuw and Leroy (1988).

The minimum covariance determinant estimator (MCD) is another method with high breakdown point (Rousseeuw 1985). It searches for a subset containing half of the data, the covariance matrix of which has the smallest determinant. Recently, it has been proved that the MCD estimator is asymptotically normal (Butler and Jhun 1990). The MCD estimator needs somewhat more computation time than does the MVE. The MCD estimator has also been computed by means of simulated annealing (R. Grübel, personal communication), but this approach takes much more computation time.

We have tried out two approximate algorithms for the MVE. The first is the *resampling algorithm* described in Rousseeuw and Leroy (1987). It is based on the idea of looking for a small number of good points, rather than for k bad points, where $k = 1, 2, 3, \ldots$. This resembles certain regression algorithms used by Rousseeuw (1984) and, independently, by Hawkins et al. (1984). We draw subsamples of $p + 1$ different observations, indexed by $J = \{i_1, \ldots, i_{p+1}\}$. The mean and covariance matrix of such a subsample are

$$T_J = \frac{1}{p+1}\sum_J \mathbf{x}_i \quad \text{and}$$

$$\mathbf{C}_J = \frac{1}{p}\sum_J (\mathbf{x}_i - T_J)^t(\mathbf{x}_i - T_J). \quad (A.7)$$

The corresponding ellipsoid should then be inflated or deflated to contain exactly h points, which amounts to computing

$$m_J^2 = \{(\mathbf{x}_i - T_J)\mathbf{C}_J^{-1}(\mathbf{x}_i - T_J)^t\}_{h:n} \quad (A.8)$$

because m_J is the right magnification factor. The squared volume of the resulting ellipsoid is proportional to $m_J^{2p}\det(\mathbf{C}_J)$, of which we keep the smallest value. For this "best" subset J we compute

$$T(\mathbf{X}) = T_J \quad \text{and} \quad \mathbf{C}(\mathbf{X}) = (\chi^2_{p,.50})^{-1}c^2_{n,p}m_J^2\mathbf{C}_J \quad (A.9)$$

as an approximation to the MVE estimator, followed by a reweighting step as in (A.5) and (A.6). The number of subsamples J depends on a probabilistic argument, because we want to be confident that we encounter enough subsamples consisting of $p + 1$ good points. Moreover, by carrying out a simulation study, we found that $c^2_{n,p} = (1 + 15/(n - p))^2$ is a reasonable small-sample correction factor. Therefore, this factor was incorporated in all of the examples of our article.

The *projection algorithm* is a variant of an algorithm of Gasko and Donoho (1982). For each point $\mathbf{x}_i$ we consider

$$u_i = \max_{\mathbf{v}} \frac{|\mathbf{x}_i\mathbf{v}^t - L(\mathbf{x}_1\mathbf{v}^t, \ldots, \mathbf{x}_n\mathbf{v}^t)|}{S(\mathbf{x}_1\mathbf{v}^t, \ldots, \mathbf{x}_n\mathbf{v}^t)} \quad (A.10)$$

where L and S are the MVE estimates in one dimension, which we compute as follows. For any set of numbers $z_1 \le z_2 \le \cdots \le z_n$ one can determine its shortest half by taking the smallest of the differences

$$z_h - z_1, z_{h+1} - z_2, \ldots, z_n - z_{n-h+1}.$$

If the smallest difference is $z_j - z_{j-h+1}$ we put L equal to the midpoint of the corresponding half,

$$L(z_1, \ldots, z_n) = (z_j + z_{j-h+1})/2 \quad (A.11)$$

and S as its length,

$$S(z_1, \ldots, z_n) = c(n)(z_j - z_{j-h+1}) \quad (A.12)$$

up to a correction factor $c(n)$, which depends on the sample size. Note that (A.10) is exactly the one-dimensional version of the robust distance RD_i of (A.4), but applied to the projections $\mathbf{x}_i\mathbf{v}^t$ of the data points $\mathbf{x}_i$ on the direction $\mathbf{v}$. As not all possible directions $\mathbf{v}$ can be tried, we have to make a selection. We take all $\mathbf{v}$ of the form $\mathbf{x}_l - M$ where $l = 1, \ldots, n$ and M is the coordinatewise median:

$$M = (\underset{j=1,\ldots,n}{\text{median}}\ \mathbf{x}_{j1}, \ldots, \underset{j=1,\ldots,n}{\text{median}}\ \mathbf{x}_{jp}).$$

In the algorithm we update an array $(u_i)_{i=1,\ldots,n}$ while l loops over $1, \ldots, n$. The final u_i are approximations of RD_i which can be plotted or used for reweighting as in (A.5) and (A.6).

Both algorithms are very approximate, but from our experience this usually does not matter much as far as the detection of outliers is concerned. The resampling algorithm is affine equivariant but not permutation invariant, because reordering the $\mathbf{x}_i$ will change the random subsamples J. On the other hand, the projection algorithm is permutation invariant because it considers all values of l, but it is not affine equivariant. Note that the projection algorithm is much faster than the resampling algorithm, especially in higher dimensions.

[*Received November 1988. Revised August 1989.*]

REFERENCES

Atkinson, A. C. (1986), "Masking Unmasked," *Biometrika*, 73, 533–541.

Beckman, R. J., and Cook, R. D. (1983), "Outlier.....s," *Technometrics*, 25, 119–163.

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

Butler, R., and Jhun, M. (1990), "Asymptotics for the Minimum Covariance Determinant Estimator," unpublished manuscript, Colorado State University, Dept. of Statistics.

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

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, London: Chapman & Hall.

Chork, C. Y. (in press), "Unmasking Multivariate Anomalous Observations in Exploration Geochemical Data From Sheeted-Vein Tin Mineralization," *Journal of Geochemical Exploration*.

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association*, 76, 354–362.

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

Gasko, M., and Donoho, D. (1982), "Influential Observation in Data Analysis," in *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 104–109.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: John Wiley.

Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics*, 16, 619–628.

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

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984), "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," *Technometrics*, 26, 197–208.

Karnel, G. (1988), "Robust Canonical Correlation and Correspondence Analysis," unpublished manuscript, Technical University, Vienna, Dept. of Statistics.

Lopuhaä, H. P., and Rousseeuw, P. J. (in press), "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices," *The Annals of Statistics*.

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

——— (1985), "Multivariate Estimation With High Breakdown Point," in *Mathematical Statistics and Applications* (Vol. B, eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht: Reidel Publishing, pp. 283–297.

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

——— (1988), "A Robust Scale Estimator Based on the Shortest Half," *Statistica Neerlandica*, 42, 103–116.

Stahel, W. A. (1981), "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen," unpublished Ph.D. thesis, ETH Zürich.

Stromberg, A. J. (1989), "Nonlinear Regression With High Breakdown Point," unpublished Ph.D. thesis, Cornell University, School of Operations Research and Industrial Engineering.