Multivariate Outliers

NICK FIELLER

Volume 3, pp. 1379–1384

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Multivariate Outliers

The presence of outliers in univariate data is readily detected because they must be extreme observations that can easily be identified numerically or graphically. The outliers are amongst the largest or smallest observations and it is not difficult to identify and examine these and perhaps test them for discordancy in relation to the model proposed for the data (*see* **Outliers**; **Outlier Detection**). By contrast, outliers in multivariate data present special difficulties, primarily because of lack of a clear definition of the 'extreme observations' in such a data set. While the intuitive notion that the extreme observations are those that are 'furthest from the main body of the data', this does not help locate them since there are many possible 'directions' in which they could be separated. As many authors have commented, univariate outliers 'stick out at one end or the other but multivariate outliers just stick out somewhere'.

Figure 1 shows a plot of two components, B and I, from a set of data containing measurements of nine trace elements (labelled A, B, ..., I) measured in a collection of 269 archaic clay vessels. There is clearly one outlier 'sticking out' in a direction of forty-five degrees from the main body of the data and there are maybe some suspicious observations sticking out in the opposite direction.

Most proposed techniques for pinpointing any outliers that may then be tested formally for discordancy (or, maybe, just examined for deeper understanding of the data) rely on calculations of sample statistics such as the mean and covariance matrix, which

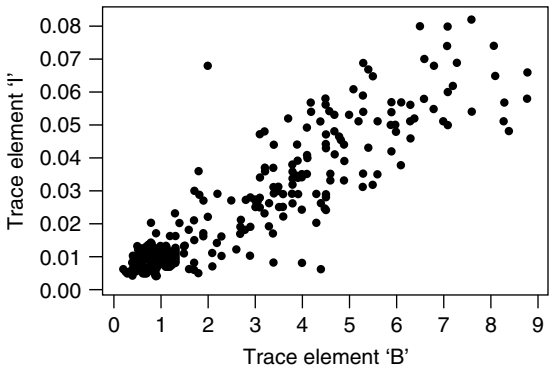

Figure 1 Outlier revealed on bivariate plot

can be seriously affected by the outliers themselves. With univariate data, inflation of sample variance and change in mean does not camouflage their hiding place; by contrast, distortion of the sample covariance can hide multivariate outliers effectively, especially when there are multiple outliers placed in different directions. This feature makes the use of robust estimates of the mean and covariance matrix a possibility to be considered, though this may or may not be effective. This is discussed further below.

As with all outlier problems in the whole range of contexts of statistical analysis, there are three distinct objectives: *identification, testing for discordancy* and *accommodation*. Irrespective of whether the first two have been considered, the third of these can be handled by the routine use of robust methods; specifically, use of robust estimates of mean and covariance (*see* **Robust Statistics for Multivariate Methods**). These methods are widely available in many statistical software packages, for example, R, S-Plus, and SAS and are not specifically discussed here (*see* **Software for Statistical Analyses**). Identification may be followed by a formal test of discordancy or it may be that the identification stage has merely revealed simple mistakes in the data recording or, maybe some unexpected feature of interest in its own right. Identification may proceed in step with formal tests in hunting for multiple outliers where *masking* and *swamping* are ever-present dangers.

Identification

Strategies for identification begin with simple graphical methods as an aid to visual inspection. For an $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ representing a random sample of n p-dimensional observations $\mathbf{x}_i$, $i = 1, \ldots, n$, univariate plots of original components (or **boxplots** or listing the ordered values numerically) should reveal any simple extreme potential recording errors. Bivariate (or pseudo three-dimensional (*see* **Three Dimensional (3D) Scatterplots**) plots of original components are only worthwhile for low dimensional data, perhaps for $p \leq 10$ or 15, say (an effective upper limit for easy scanning of matrix plots of all pairwise combinations of components). These may not reveal anything other than simple recording errors in a single component which are not extreme enough to have caused concern on first examination of just that component alone. Further, they will not

reveal outliers that are not simple recording errors but are attributable to more interesting unexpected causes. Figure 1 above shows just such an example where the outlier could be a low value of element B or a high value of element I. They are nevertheless worthwhile, especially since such matrix plots are quick and easy to produce (*see* **Scatterplot Matrices**). For higher dimensional data sets, it is not easy to focus on all of the bivariate scatterplots individually.

A more sophisticated approach is to use an initial dimensionality reduction method as a preliminary step to reduce the combinations of components that need to be examined. The obvious candidate is **principal component analysis**, based on either the covariance or the correlation matrix (or on both in turn) and examining bivariate scatterplots of successive pairs of principal components. These methods have the key advantage in that they can be used even if $n < p$, that is, more dimensions than data points. Here, the rationale is that the presence of outliers will distort the covariance matrix resulting in 'biasing' a principal component by 'pulling it towards' the outlier, thus allowing it to be revealed as an extreme observation on a scatterplot including that principal component. Clearly, a gross outlier will be revealed on the high order principal components while a minor one will be revealed on those associated with the smaller **eigenvalues**. It is particularly sensible to examine plots of PCs around the 'cutoff point' on a scree plot of the eigenvalues since outliers might increase the apparent effective dimensionality of the data set and so be revealed in this 'additional' dimension. These methods can be effective for either low numbers of outliers in relation to the sample size or for 'clusters' of outliers that arise from a common cause and so 'stick out' in the same direction. For large numbers of heterogeneous outliers, then, a robust version of principal component analysis is a possibility, although this loses the rationale that the outliers distort the PCs towards them by distorting the covariance or correlation matrix.

Returning to the measurements of nine trace elements of the clay vessels, Figure 2 shows a scree plot of cumulative variance explained by successive principal components calculated from the correlation matrix of the nine variables on a subset of 53 of the vessels. This suggests that the inherent dimensionality of the data is around five, with the first five principal components containing 93% of the variation and so it could be that outliers would be revealed on plots

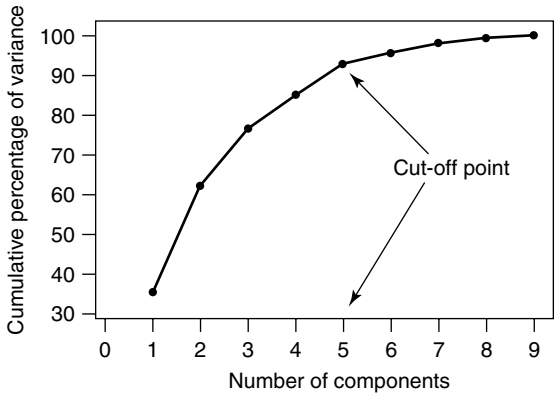

Figure 2 Scree plot of cumulative percentage of variance

of principal components around that value. Figure 3 below shows scatter plots of the data referred to successive pairs of the first five principal components.

Inspection of these suggests that the two observations on the right of the lower right-hand plot (i.e. with high scores on the fifth component) would be worth investigating further, though they are not exceptionally unusual. Closer inspection (with interactive brushing) reveals that one of this pair occurs as the extreme top right observation in the top left plot of PC1 versus PC2 in Figure 3, that is, it has almost the highest scores on both of the first two principal components. Further inspection with brushing reveals that the group of three observations on the top right of PC3 versus PC4 (i.e. those with the three highest scores on the fourth component) have very similar values on all other components and so might be considered a group of outliers from a common source.

A more intensive graphical procedure, which is only viable for reasonably modest data sets (say $n <\sim 50$) with more observations than dimensions ($n > p$) and where only a small number of outliers is envisaged is the use of outlier displaying components (ODCs). These are really just linear discriminant coordinates (sometimes called *canonical variate coordinates*) (*see* **Discriminant Analysis**). For a single outlier, these ODCs are calculated by dividing the data set into two groups, one consisting of just one observation $\mathbf{x}_j$ and the other of the $(n-1)$ observations omitting $\mathbf{x}_j$, $j = 1, \ldots, n$ (so that there is potentially a separate ODC for each observation). It is easily shown that the univariate likelihood ratio test statistic for discordancy of x_j under a Normal mean slippage model calculated from the values projected

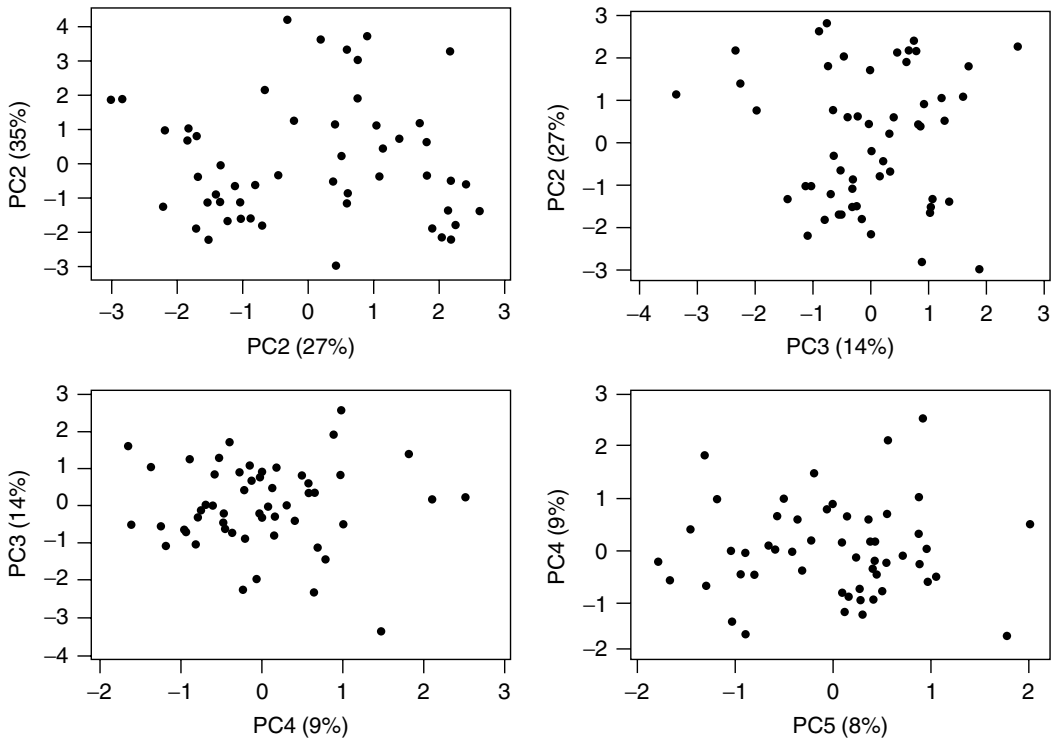

Figure 3 Pairwise plots on first five principal components

onto this ODC is numerically identical to the equivalent statistic (see below), calculated from the original p-dimensional data. Thus, the one-dimensional data projected onto the ODC capture all the numerical information on the discordancy of that observation (though the reference distribution for a formal test of discordancy does depend on the dimensionality p). Picking that with the highest value of the statistic will reveal the most extreme outlier. The generalization to two outliers and beyond depends on whether they arise from a common or two distinct slippages. If the former, then there is just one ODC for each possible pair and, if the latter, then there are two. For three or more, the number of possibilities increases rapidly and so the use is limited as a tool for pure detection of outliers to modest data sets with low contamination rates. However, the procedure can be effective for displaying outliers already identified and for plotting subsidiary data in the same coordinate system. Examination of loadings of original components may give information on the nature of the outliers in the spirit of union-intersection test procedures.

Returning to the data on trace elements in claypots, the plot in Figure 4 on the left displays the identified outlier in Figure 1 on the single outlier displaying component (ODC) for that observation as the horizontal axis. This has been calculated using all nine dimensions rather than just two as in Figure 1 and this contains all the information on the deviation of this observation from the rest of the data. The vertical axis has been chosen as the first principal component but other choices might be sensible, for example, a component around the cutoff value of four or five or else on a 'subprincipal component': that vector which maximizes the variance subject to the constraint of being orthogonal with the ODC. Examination of the loadings of the trace elements in the ODC show heavy (negative) weighting on element I with moderate contributions from trace elements F (negative) and E (positive). Trace element B, the horizontal axis in Figure 1, has a very small contribution and, thus, it might be suspected that any recording error is in the measurement of element I. A matrix plot of all components (not shown here) does not

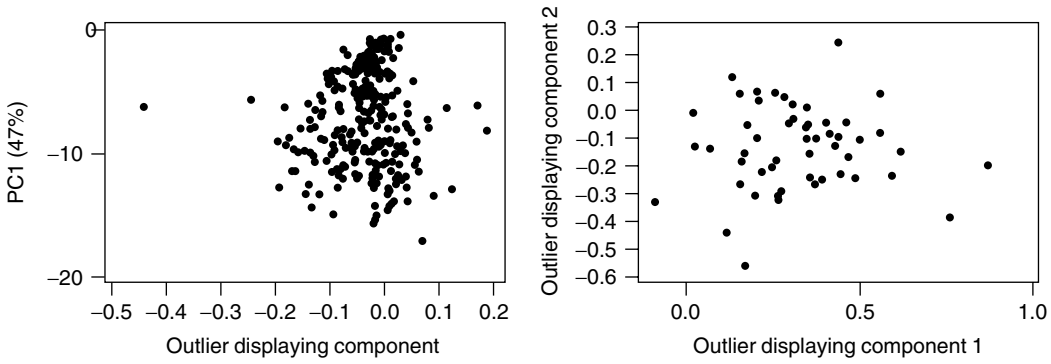

Figure 4 Illustrations of outlier displaying components

reveal this, primarily because this pair of elements is the only one exhibiting any strong correlation. It may be seen that there are several further observations separated from the main body of the data in the direction of the ODC, both with low values (as with the outlier) and with high values. Attention might be directed towards these samples. It should be noted, of course, that the sign of the ODC is arbitrary in the sense that all the coefficients could be multiplied by -1 (thus reflecting the plot about a vertical axis), in the same way that signs of principal components are arbitrary.

The figure on the right of Figure 4 displays the subset of the data discussed in relation to Figure 3. Two groups of outliers have been identified. One is the pair of extreme observation on the fifth principal component and these appear on the right of the plot. The other is the triple of extreme observations on the fourth principal component. These appear in the lower left of the plot. However, two further observations appear separated from the main body in this coordinate system (at the top of the plot) and further examination of these samples might be of interest. Informal examination of the loadings of these two ODCs suggests that the first is dominated by trace elements D and E and the second by E, F and I. Whether this information is of key interest and an aid in further understanding is, of course, a matter for the scientist involved rather than purely a statistical result.

The methods outlined above have the advantage that they display outliers in relation to the original data – they are methods for selecting interesting views of the raw data that highlight the outliers. For data sets with large numbers of observations and dimensions, this approach may not be viable and various, more systematic approaches have been suggested based on some variant of probability plotting. The starting point is to consider ordered values of a measure of generalized squared distance of each observation from a measure of central location $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma}) = (\mathbf{x}_j - \mathbf{C})'\mathbf{\Gamma}^{-1}(\mathbf{x}_j - \mathbf{C})$, where $\mathbf{C}$ is a measure of location (e.g. the sample mean or a robust estimate of the population mean) and $\mathbf{\Gamma}$ is a measure of scale (e.g., the sample covariance or a robust estimate of the population value). The choice of the sample mean and variance for $\mathbf{C}$ and $\mathbf{\Gamma}$ yields the squared **Mahalanobis distance** of each observation from the mean. It is known that if the $\mathbf{x}_j$ comes from a multivariate Normal distribution (*see* **Catalogue of Probability Density Functions**), then the distribution of the $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma})$ is well approximated by a gamma distribution (*see* **Catalogue of Probability Density Functions**) with, dependent upon $\mathbf{\Gamma}$, some shape parameter that needs to be estimated. A plot of the ordered values against expected order statistics would reveal outliers as deviating from the straight line at the upper end. Barnett and Lewis [1] give further details and references. It should be noted that, typically, these methods are only available when $n > p$; in other cases, a (very robust) dimensionality reduction procedure might be considered first.

Further, these methods are primarily of use when the number of outliers is relatively small. In cases where there is a high multiplicity of outliers, the phenomena of *masking* and *swamping* make them less effective. An alternative approach, typified by Hadi [2, 3], is based upon finding a 'clean' interior subset of observations and successively adding more observations to the subset. Hadi recommends starting

with the $p + 1$ observations with smallest values of $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma})$, with very robust choices for $\mathbf{C}$ and $\mathbf{\Gamma}$. An alternative might be to start by *peeling* away successive *convex hulls* to leave a small interior subset. Successive observations are added to the subset by choosing that with the smallest value of $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma})$, calculated just from the clean initial subset (with $\mathbf{C}$ and $\mathbf{\Gamma}$ chosen as the sample mean and covariance) and with the ordering updated at each step. Hadi gives stopping criterion based on simulations from multivariate Normal distributions, the remaining observations not included in the clean subset being declared discordant.

Tests for Discordancy

Formal tests of discordancy rely on distributional assumptions; an outlier can only be tested for discordancy in relation to a formal statistical model. Most available formal tests presume multivariate Normality; exceptions are tests for single outliers in bivariate exponential and bivariate Pareto distributions. Barnett and Lewis [1] provide some details and tables of critical values for these two distributions. For multivariate Normal distributions (*see* **Catalogue of Probability Density Functions**, the likelihood ratio statistic for a test of discordancy of a single outlier arising from a similar distribution with a change in mean (i.e., a mean slippage model) is easily shown to be equivalent to the Mahalanobis distance from the sample mean (i.e., $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma})$, with $\mathbf{C}$ and $\mathbf{\Gamma}$ taken as the sample mean and covariance). In one dimension, this is equivalent to the studentized distance from the sample mean. Barnett & Lewis [1] give tables of 5% and 1% critical values of this statistic for various values of $n \leq 500$ and $p \leq 5$, together with similar tables for critical values of equivalent statistics when the population covariance and also both population mean and covariance are known, as well as the case where an independent external estimate of the covariance is available. Additionally, they provide similar tables for the appropriate statistic for a pair of outliers and references to further sources.

Implementation

Although the techniques described here are not specifically available as an 'outlier detection and testing module' in any standard package, they can be readily implemented using standard modules provided for principal component analysis, linear discriminant analysis and robust calculation of mean and covariance, provided there is also the capability of direct evaluation of algebraic expressions involving products of matrices and so on. Such facilities are certainly available in R, S-PLUS and SAS. For example, to obtain a single outlier ODC plot, a group indicator needs to be created, which labels the single observation as group 1 and the remaining observations as group 2 and then a standard linear discriminant analysis can be performed. Some packages may fail to perform standard linear discriminant analysis if there is only one observation in the group. However, it is easy to calculate the single outlier ODC for observation $\mathbf{x}_j$ directly as $\mathbf{\Gamma}^{-1}\mathbf{x}_j$, where $\mathbf{\Gamma}$ is the sample covariance matrix or a robust version of it. The matrix calculation facilities would probably be needed for implementation of the robust versions of the techniques.

References

[1] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Edition, Wiley, New York.

[2] Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society Series B* **54**, 761–771.

[3] Hadi, A.S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal. Statistical Society Series B* **56**, 393–396.

NICK FIELLER