MARDI Res. Bull., (1982) 10, 1: (43—51)

OUTLIERS IN MULTIVARIATE DATA EXPOSURE THROUGH THE GRAPH-THEORETICAL METHODS

MD. KHIR ABDUL RAHMAN*

Keywords: Outliers, Data exposure, Multidimensional plot, Minimally connected graphs, Normal quantile, Edge lengths.

RINGKASAN

Dalam melakukan analisa data, selain dari mengemukakan ringkasan atau penemuan akhir untuk data tersebut, pertimbangan harus juga diberi kepada aspek penapisan data. Yang utama ialah proses pendedahan data tersebut untuk mengesan kehadiran titik-titik data pencilan. Masalah pendedahan titik data pencilan dari data multivariate dibincang di sini menurut pendekatan secara graf-teoretik. Dua kaedah dikajikan: yang pertama melibatkan teknik memplot data multidimensi dan yang kedua menggunakan prosedur pemangkasan ke atas graf terkait minimum.

INTRODUCTION

The recent development in biometry has been the emphasis on the analysis of actual research data as contrasted with the computations of statistics that are simply abstractions or summaries from these data. This area is known as data analysis – defined as the systematic search of sets of data to yield both summaries of and fresh insights into the relationship in a given study.

TUKEY and WILK (1966) broadly outlined data analysis through the following aspects:

(ii) Summarization – the presentation of results in terms of statistics (such as mean, standard deviation, correlation, etc.)

(i) Exposure – the presentation of analyses so as to facilitate the detection of not only anticipated but also unanticipated characteristics of data, and

Statistical methodology of summarization has often been developed without specific attention to the exposure value of the techniques. The goals of summarization has always been artificially separated from the objective of exposure, though both can be viewed as the prerequisites to data analysis.

This paper touches on one technique of exposure in data analysis, namely the detection of outliers. Particular emphasis will be given to the use of graph-theoretical methods as applied to multivariate data.

In a sample of n observations, it is possible that a limited number of them are so far separated in value from the remainder that they give rise to the question whether they are not from a different population, or that the sampling technique is at fault. Such observations are called **outliers.**

The occurrences of outliers are common in research data. In the study on nutrient requirement of English cabbage (PURUSHOTHAMAN and JOSEPH, 1977), for example, the presence of outliers may be suspected in some of the scatter plots given. A genuine outlying point, if considered, would significantly alter the regression relationship thus computed. The same may be said about trend determination in economic analysis, where random and outlying points are common.

Thus, quite often, an analyst is faced with this problem of deciding whether certain of his observations properly belong in his presentation of measurements obtained. He must decide whether these observations are valid. Two objectives may be given to this problem of outliers detection (QUEENSBERRY and DAVID, 1961). One might, for instance, be primarily interested in prunning the observations in order to secure a more accurate analysis of outliers.

*Computer Centre, MARDI, Serdang.

Or, one might be particularly interested in identifying which are the genuinely exceptional observations in order to create a new insight into the phenomena under study.

In this paper, the analytic data philosophy with regard to detection of outliers is more of the second case. Outlier identification process here deals primarily with the problem of isolating peculiarities, not thereby implying or advising rejection of outliers – in essence, it is just a tool of data analysis.

DIXON (1953) characterises outliers in univariate case according to the following model.

Let N (μ,σ^2) represents a population with mean μ and variance σ^2. An observation from N $(\mu+\lambda\sigma,\sigma^2)$ introduced into a sample from N (μ,σ^2) is termed a **location error.** An observation from N $(\mu,\lambda^2\sigma^2)$ introduced into a sample from N (μ,σ^2) is a **scalar error**.

In such a univariate case, the use of trimmed and Winsorised means for the detection of outliers has been advocated by DIXON (1960) and by TUKEY (1960; 1962).

In the treatment of outlier detection considered in this study, the outlying observations will be assumed to be from a k-variate normal distribution with shifted mean vector, specifically from N $(\mu+\alpha,\Sigma)$. The remaining observations are from a N (μ, Σ) distribution. The case of k=2 and k=3 will be considered for illustration.

THE EFFECT OF DATA CONTAMINATION

The consequences of having defective observations in a multivariate sample are intrinsically more complex than in the univariate case. One reason for this is that multivariate outliers can distort not only measures of location and dispersion, but also those of orientation e.g. correlation (see GNANADISEKAN and KETTENRING, 1972, for details).

The effect of the presence of these outliers will be demonstrated using 51 sample points artificially generated from a bivariate normal distribution with zero mean vector, unit variances and a correlation, $\rho = 0.7$. Additive disturbance of two observations will simulate the presence of outliers.

Table 1 shows the effect of two outliers on the means, variances and correlation for the bivariate case. It is evident from the table that the presence of only two outliers effectively influences the sample estimates, especially the means and the variances. Though only at two-percent shift in the correlation is observed here, DEVLIN *et al.*, (1975), however reported that the presence of an outlier in a generated bivariate normal distribution of size 60 with $\rho = 0.9$ has the effect of shifting the correlation by an average of 8%!. It is thus necessary, for this reason alone, to identify the outlying observations so that a better and correct analysis of the data is possible.

DETECTION OF MULTIVARIATE OUTLIERS

A brief overview of several outlier detection techniques will be given here. Subsequently, elaborations over some graphical techniques will be made to demonstrate their effectiveness in detecting two types of outliers.

In the univariate case, the outliers are simply the largest and the smallest observations. In DIXON's (1950) gap test, one tests whether the length of an interval between the suspected outlier and its closest or second closest variate is too large in comparison to the observed range of variability. DOORNBOS (1958) uses the same interval length between the outlier and its closest variate, but instead of the observed range of variability, the standard deviation of the good observations (without outlier) is used for comparison.

Detection of outliers in samples from k-variate normal distribution is much more

TABLE 1: THE EFFECT OF DATA CONTAMINATION

Statistics	Without outlier	With 2 outliers	Deviation in estimates
$\bar{V}_1$	0.0276	0.0409	0.0433 (48%)
$\bar{V}_2$	0.1173	0.1333	0.0160 (13%)
Var (V_1)	0.8927	1.0360	0.1433 (16%)
Var (V_2)	0.7733	0.9016	0.1283 (17%)
Correlation	0.7103	0.6946	0.0157 (2%)

difficult than in the corresponding univariate problem since there is no quarantee that an outlier can be detected visually either by examining each variable separately or with bivariate scatter diagrams. The complexity of the problem is further compounded by the fact that there exist outliers which distort the estimates for correlation without affecting the variances. Also there are outliers which behave in the contrary.

Despite the apparent complexity of the problem, multivariate outliers can still be detected by the fact that they are somewhat isolated from the main cloud of points. WILKS (1963) gives an identification procedure based on outliers scatter ratios. For the ℓ-th observations, the 'one outlier scatter ratio' is

$$S_\ell = |A_{ij\ell}| / |A_{ij}|$$

where $|A_{ij}|$ is the 'internal scatter' and $|A_{ij\ell}|$ is the analogous quantity with $\underline{x}_\ell$ ommitted.

GUTTMAN (1973) presents a Bayesian approach to the identification of a single outlier in a multivariate normal distribution. He assumes that all but one observation comes from a N (μ, Σ) distribution, the anomalous observation comes from N ($\mu+\alpha$, Σ). Outlier identification is based on the posterior distribution of α. HAWKINS (1976) and FELLEGI (1975) based their multivariate outlier procedures on the ordering of the principal components while KARLIN and TRUAX (1969) discuss tests of multivariate outliers using studentized distance measures and a slippage-type alternative hypothesis.

However, there exists a class of graphical and graph-theoretical techniques which can be implemented for detecting outliers. The following section will treat two of those techniques with the aim of illustrating their efficiencies, considered in terms of the ease in implementation and negligible analysis effort. The two techniques are namely (i) plot of high dimensional data and (ii) analysis of a minimally-connected graph.

GRAPHICAL METHODS

To illustrate outlier detection using graphical methods, the artificially generated data as mentioned earlier will be used. The 51 observations, are from a tri-variate distribution with zero mean vector, unit variances and a positive Σ. *Figure 1* shows the bivariate scatter diagram with two observations shifted to simulate the presence of outliers. The outlier X_1 may be detected by simply using the univariate test while X_2 is not detectable with a univariate test.

1. Plot of k-dimensional data

ANDREWS (1972) proposed a very simple plotting procedure. If the data is k-dimensional, each point $\underline{x} = (x_1, x_2, \ldots, x_k)^T$ defines a function

$$f_{\underline{x}}(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + \ldots.$$

where the function values $f_{\underline{x}}(t)$ are projections of the vector $\underline{x}$ on the vector

$$f_t(t) = (1/\sqrt{2}, \sin t, \cos t, \sin 2t, \cos 2t, \ldots).$$

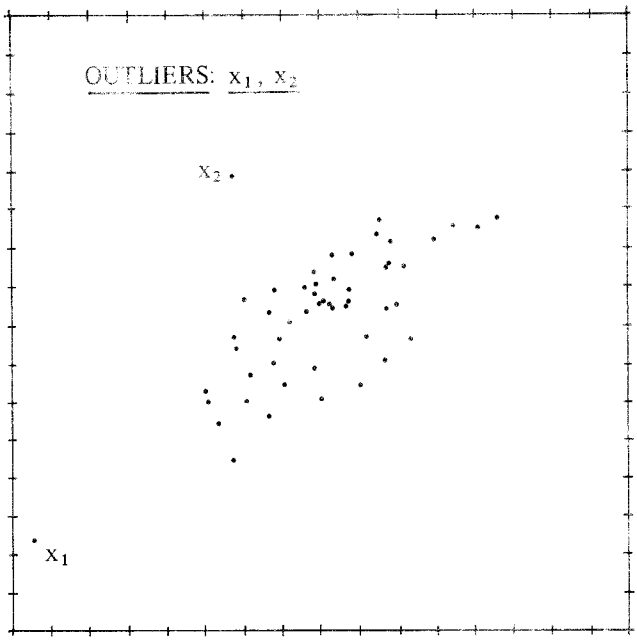

Figure 1. Bivariate scatter of 51 points

When normalised, this vector defines a point on the k-sphere and as t varies, this point describes a periodic curve on the k-sphere. The function is plotted over the range $-\pi < t < \pi$.

Several useful properties are attributed to this function, namely

(i) The function preserves means. If $\bar{\underline{X}}$ is the mean vector of set of n observations, then

$$f_{\bar{\underline{X}}}(t) = \frac{1}{n} \sum_{i=1}^{n} f_{\underline{X}_i}(t)$$

(ii) The function preserves variance. If the components of the data are uncorrelated with common variance σ^2, then

$$\text{Var}\,[f_{\underline{X}}(t)] = \sigma^2\,[\tfrac{1}{2} + \sin^2 t + \cos^2 t + \ldots.]$$

(iii) The function representation preserves distances, that is for observation $\underline{x}_1$, and $\underline{x}_2$, the familiar Euclidean distance is

$$\| f_{\underline{X}_1}(t) - f_{\underline{X}_2}(t) \|_{L_2} = \pi \sum_{i=1}^{k} (x_{1i} - x_{2i})^2$$

This third property is the basis for outlier detection. Because of this relation, close points will appear as close functions and distant points as distant functions. Thus, multivariate outliers may be identified visually from the plot of the functions.

Figure 2 shows the plot of $f_{\underline{X}}(t)$ for 17 three-dimensional observations from the artificially generated data set. Two simulated outliers are included. It is clearly evident from the figure that the periodic curves of the two outliers are well identified. For outlier x_1, the distinct behaviour of the curve may be observed in the whole range of t, while for outlier x_2 the distinctness is apparent in the range of t from $-\pi$ to 0.5. The remaining 15 observations define nearly similar periodic curves.

2. Analysis of minimally connected graphs

This technique makes use of the concept of adjacency between points, in essence, the distance between a point and its nearest neighbour. A convenient way to realise this is to use the edge connections of minimally connected graphs, which are graphs of n points and (n−1) edges, where the total length of the edges is minimum. The distance

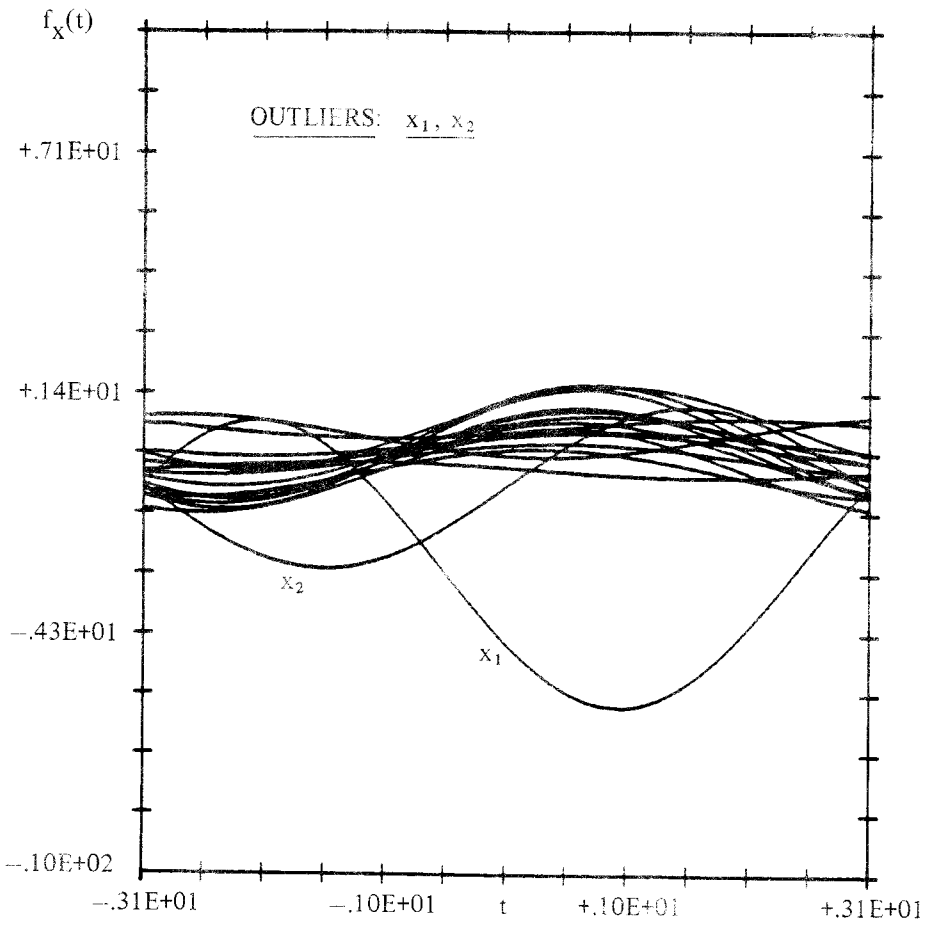

Figure 2. $f_X(t)$ Plot of 17 Trivariate Points having 2 Outliers

between two adjacent points is thus the length of the edge connection of the two points of the MC graph. In such a graph, an outlier may be characterised as the point whose edge connection is relatively long. Therefore, in detecting outliers, all that is needed is the information on the (n−1) distances of the graph. The MC graph can be generated using the distance function,

$$d_{ij} = [\sum_{\ell=1}^{n-1} (x_{\ell i} - x_{\ell j})^2]^{1/2}$$

as a measure of proximity between any pair of points i and j. Algorithms for constructing MC graphs are given by PRIM (1957), KRUSKAL, (1956) and WHITNEY (1972), among others. *Fig. 3* below shows the MC graph constructed for the scatter of points of *fig. 1*, using the distance function above. The two simulated outliers are indicate...

2.1 Normal quantile plots

Let d_i (i = 1, ..., m) denote the lengths of the m = n−1 edges of the MC graphs. If normal distribution is assumed for d_i and d_i^2 then the normal plots of both the variable would reveal a linear pattern. Any outliers present would be indicated by deviations from linearity. *Figure 4* shows the quantile plot of d_i and d_i^2 for the bivariate and trivariate case. In all the four cases, the simulated outliers are clearly indicated, two for the bivariate and three for the trivariate sample. The presence of linearity when d_i^2 is used, and not when d_i values are plotted, indicates that the assumption of normal distribution is more appropriate for d_i^2. GNANADESIKAN and KETTENRING (1972) however recommend the assumption of gamma distribution for the d_i^2 values.

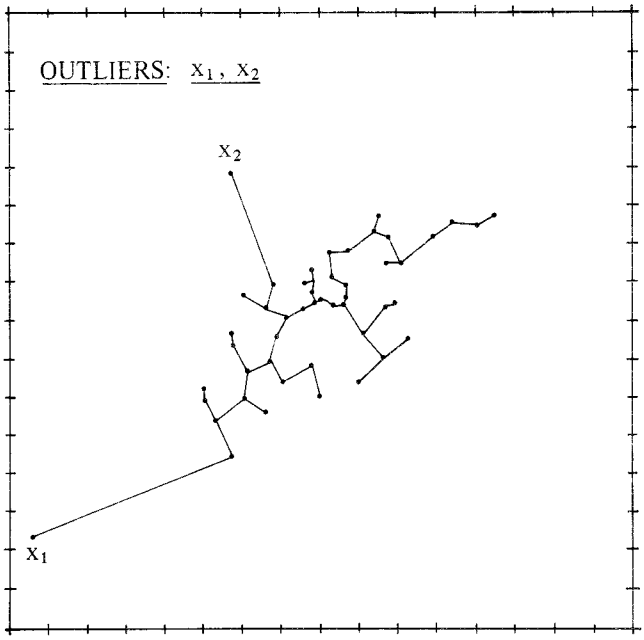

Figure 3: Minimally connected graph of 51 points (k = 2).

2.2 Trimming the MC graphs for outliers

This technique is based on the fact that a multivariate outlier must 'stick out' somewhere from the hyper-cloud of points. On an MC graph, this idea is manifested in the presence of edges which project themselves away from the central body of points.

Such edges, d_i, may be detected using the following criteria:-

1. $d_i > f\,\bar{d}_i$

2. $d_i > \bar{d}_i + ts_i$

where $\bar{d}_i$ and s_i are mean and standard deviation of h neighbouring edge lengths. f and t are multiplying factors.

Table 2 shows the number of maverick observations apparently detected using different depth factor, h. The bivariate and trivariate samples contain two and three simulated outliers, respectively. A reasonable choice of f=2, and t=3 will be used (ZAHN, 1971). The table indicates that for the two examples, a minimum depth of 20 neighbouring edges is sufficient in order to be able to detect the simulated outliers. This corresponds to a forty percent depth factor. Also, by using this depth factor, it was found that outliers were detected correctly in the following range of f and t

bivariate	$1.5 < t < 3.0$; $0.0 < f < 3.5$
	$0.0 < t < 1.5$; $2.0 < f < 3.5$
trivariate	$1.0 < t < 2.5$; $0.0 < f < 3.5$
	$0.0 < t < 1.0$; $2.0 < f < 3.5$

CONCLUSION

The graphical techniques of detecting outliers have been illustrated for the 2– and 3–variable case. However, extension to the k–variable case is a trivial process.

In the k–dimensional plot, only a limited amount of information may be absorbed from one plot, and 10 observations per plot is suggested if detailed examination

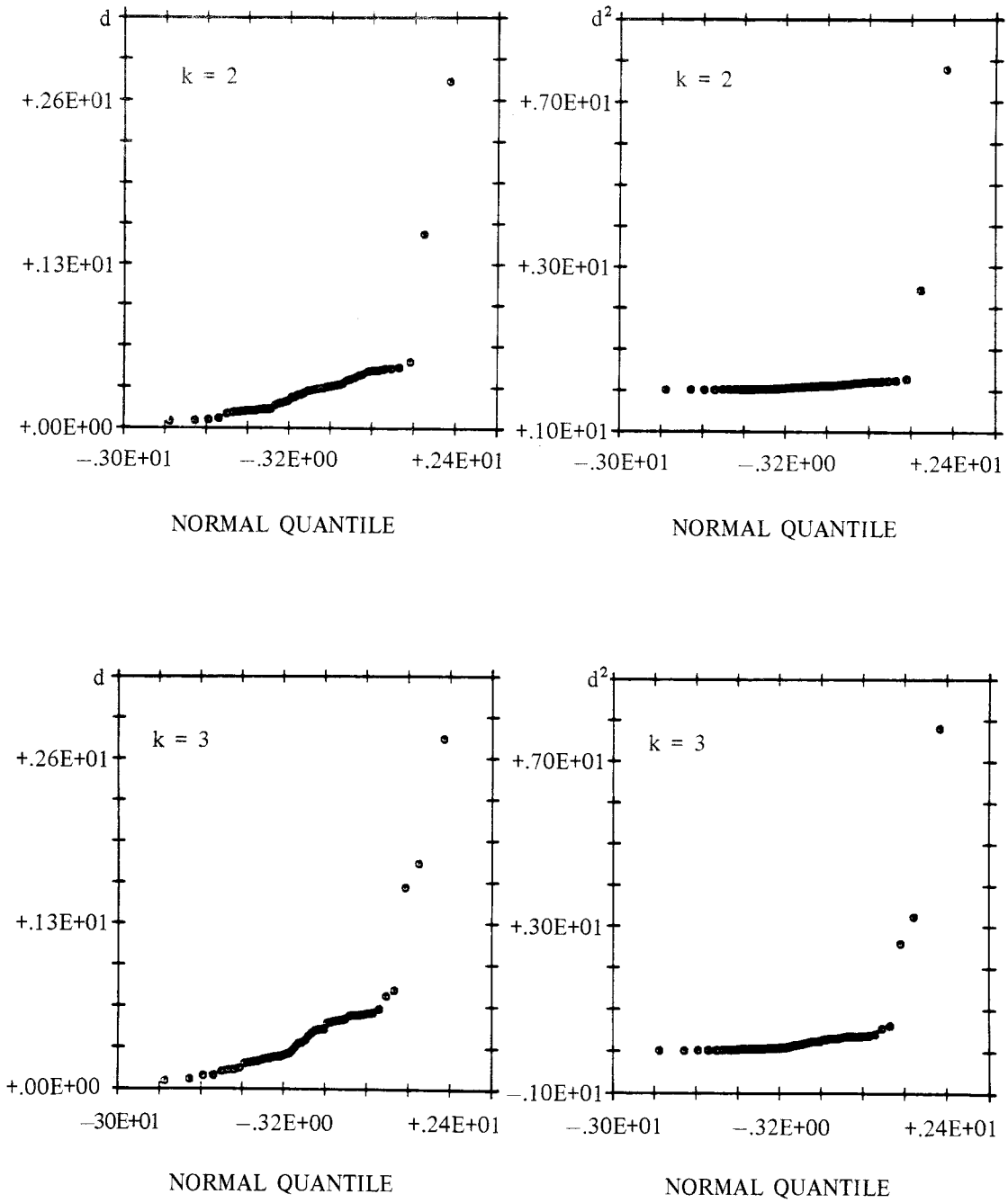

Figure 4: Normal quantile plots of d_i and d_i^2 Bivariate (k = 2), Trivariate (k = 3).

TABLE 2: NUMBER OF APPARENT OUTLIERS DETECTED FOR VARIOUS h (n = 51)

h	Number of outliers	
	Bivariate (with 2 outliers)	Trivariate (with 2 outliers)
5	3	2
10	1	2
15	2	2
20 – 50	2	3

is necessary. However, an initial plot of all observations would be a good procedure to extract general characteristics of the sample, and subsequently, separate plots of 10 points could be used to assess individual points in relation to the whole.

The ease in implementation of these graphical techniques is obvious. For the k–dimensional plots and the normal quantile plots, only a fast plotter is required. The programming effort is negligible.

The trimming of minimally connected graph implementing PAGE's (1974) algorithm is a relatively fast procedure. On the Seimens 4004 facilities, the detection of two/three outliers for a sample of size 50 takes only an average of 0.22 seconds. Normally, the computer time increase like n, but in the worst case, it may be like n^2.

ABSTRACT

In data analysis, due consideration should be given to the aspect of exposure as a supplement to the conventional data summarization. This is particularly so in exposing the presence of outliers in the data. A graph-theoretical approach is discussed in this expository problem of outlier detection for multivariate data. Two methods are reviewed; the first involves a multidimensional plotting technique and the second uses a trimming procedure on minimally connected graphs.

REFERENCES

ANDREWS, D.F. (1972). Plots of high dimensional data. Biometrics, 28: 125–136.

DEVLIN, S.J., GNANADESIKAN, R. and KETTENRING, J.R. (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika, 62: 531–546.

DIXON, W.J. (1950). Analysis of extreme values. Ann. Maths. Statist. 21: 488–506.

DIXON W.J. (1953). Processing data for outliers. Biometrics, 9: 75–89.

DIXON, W.J. (1960). Simplified estimation from censored normal samples. Ann. Math. Statist. 31: 385–391.

DOORNBOS, R. (1958). On slippage tests. Indag. Math. 20: 38–55.

FELLEGI, F.P. (1975). Automatic editing and imputation of quantitative data. I.S.I. conference, Warsaw.

FERGUSON, T.S. (1961). On the rejection of outliers. Proc. 4th. Berkeley Symp. Math. Stat. Prob. 1: 253–287.

GNANADESIKAN, R. and KETTENRING, J.R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. Biometrics, [illegible]–124.

GUTTMAN, I. (1973). Care and handling of univariate or multivariate outliers in detecting spuriosity – a Bayesian approach. Technometrics, 15: 723–738.

HAWKINS, D.M. (1974). The detection of errors in multivariate data using principal components. J. Amer. Statist. Ass., 69: 340–344.

KARLIN, S. and TRAUX D. (1960). Slippage problems. Ann. Math. Statist., 31: 296–324.

KRUSKAL, J.B. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. Proc. Amer. Math. Soc., 7: 48–50.

PAGE, R.L. (1974). A minimal spanning tree clustering method. Comm. of the ACM, 17(6): 321–323.

PRIM, R.C. (1975). Shortest connection networks and some generalisations. Bell. System Tech. Jour. 36: 1389–1401.

PURUSHOTHAMAN, V. and JOSEPH, K.T. (1977). Major nutrient requirement of English cabbage (*Brassica oleracea* var. *capitata*). MARDI Res. Bull. 5(2): 47–55.

TUKEY, W.J. (1960). A survey of sampling from contaminated distributions. In: Contribution to Probability and Statistics, Olkin, I. (Ed.), Stanford Univ. Press, p. 448–485.

TUKEY, W.J. (1962). The future of data analysis. Ann. Math. Statist. 33: 1–67.

TUKEY, W.J. and WILK, M.B. (1966). Data analysis and statistics; An expository overview. AFIPS Conf. Proc., Fall Joint Comp. Conf. 29: 695–709.

QUEENSBERRY, C.P. and DAVID, N.A. (1961). Some tests for outliers. Biometrika, 48: 379–387.

WHITNEY, V.K.M. (1972). MST. Comm. of the ACM, 15: 273–274.

WILKS, S.S. (1963). Multivariate Statistical outliers. Sankya Ser. A, 25: 407–426.

ZAHN, C.T. (1971). Graph theoretical methods for detecting and describing gestalt clusters. IEEE Trans on Computers C–20, p. 68–86.