e-ISSN 2320 –7876 www.ijfans.org
Vol.11, Iss.9, Dec 2022
© 2012 IJFANS. All Rights Reserved

MULTIVARIATE ANALYSIS: DETECTION OF MULTIPLE OUTLIERS MISSING DATA

JAMES LOURDRAJ
Research Scholar
M.Phil Mathematics
Bharath Institute Of Higher Education And Research
Mail Id : anbu.cs10@gmail.com

Dr. D. VENKATESAN
Assistant Professor, Department Of Mathematics
Bharath Institute Of Higher Education And Research
Address for Correspondence

JAMES LOURDRAJ
Research Scholar
M.Phil Mathematics
Bharath Institute Of Higher Education And Research
Mail Id : anbu.cs10@gmail.com

Abstracct

For many years, statisticians have been interested in locating "outlying," "unusual," or "unrepresentative" observations as a prelude to data analysis. Data that has been entered improperly or that does not belong to the population from which the rest of the data was collected may cause estimates to be skewed and findings to be misleading. In a number of circumstances, methods have been developed to detect and/or accommodate outlier findings. Scientists are gathering huge data sets thanks to recent technological advancements, and analysts are delving deeper to uncover the secrets of data. As a result, having a solid technique in place for dealing with rogue findings that may go unnoticed in a normal data analysis is critical.

Introduction

Consider a scientist researching a certain mosquito species. He would not be interested in the features of other kinds of mosquitoes in his data collection; instead, he would simply wish to delete the observations or verify that the observations did not affect the statistical estimations of the original population. The methods should handle the outliers in this scenario, but they do not need to identify and reject them in the estimate, and are therefore referred to be robust. As a result, robustness denotes a lack of susceptibility to minor departures from the assumptions (Huber, 1981).

It's essential to have some understanding of why or how outliers developed, in addition to recognising or tolerating them. The kinds of variation are divided into three categories by Barnett and Lewis (1994).

Univariate Outliers

The notion of outlier seems to be very easy to describe in univariate data. Outliers are data points that are "far apart" from the bulk of the data and "likely do not fit the model." A basic data plot, such as a scatter plot, stem-and-leaf plot, QQ-plot, or other similar layout, may frequently show which points are outliers. Because it strikes between the eyes, this is often referred to as the "interoccular test."

Tukey (1977) popularised the boxplot as a graphical method for identifying outliers in univariate data. If observations fall beyond the interval, the boxplot rule classifies them as outliers.

$$(Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)) \qquad (1.1)$$

The ith quartile is denoted by Q. The most frequent values for k are 1.5 for "out" values and 3.0 for "far out" observations. The chance of declaring outliers when none exist varies with the amount of observations since this criterion is not sample-size dependent. In this way, it varies from conventional outlier detection methods, which are based on the likelihood of detecting outliers when none exist.

The popular boxplot outlier labelling criterion, according to Hoaglin et al. (1986), is very permissive, with a 50% probability of identifying at least one outlier given data from a random normal sample of size 75. The rule was updated by Hoaglin and Iglewicz (1987) to make it sample-size dependent, such that the probability remains at 5% for normal samples up to 300 observations. This modified approach was extended by Banerjee and Iglewicz (2007) to handle large sample situations and a wide range of continuous univariate distributions. Kimber (1990) changed the usual boxplot outlier-labeling method for skewed distributions somewhat by substituting

$$Q_3+k(Q_3-Q_1) \text{ by } Q_3+k(Q_3-M) \text{ and } Q_1-k(Q_3-Q_1) \text{ by } Q_1-k(M-Q_1) \qquad (1.2)$$

M stands for the sample median. Kimber also used k = 1.5 to investigate the exponential distribution, including right-censored data, and utilised the Kaplan-Meier estimator to get the median and quartiles for censored data. Two univariate outlier identification techniques were proposed by van der Loo (2010). The majority of observed data is approximated in both approaches by regression of observed values on their predicted QQ - plot locations using a model cumulative distribution function..

Multivariate Outliers

Multivariate outliers offer a greater difficulty than univariate data because visual identification of multivariate outliers is almost impossible since outliers do not "pop out" at the conclusion of the data (Gnanadesikan and Kettenring, 1972). It won't assist to depict the data in bivariate form with a systematic rotation of coordinate pairs. Several important ideas provided by Barnett and Lewis (1994) and Beckman and Cook (1983) indicate to the importance of multivariate outlier identification techniques for anomaly detection.

The breakdown point is a useful metric for describing the robustness of estimators in the face of outliers. The breakdown point of an estimator, according to Hodges (1967) and Hampel (1968, 1971), is the percentage of arbitrary contaminated data that may be given in a sample before the estimator's value becomes arbitrarily high. For location and covariance estimators, Lopuhaä and Rousseeuw (1991) provided more precise definitions of the breakdown point. The breakdown point a n* (, X) is defined for a location estimator at a collection of observations X.

as:

$$\varepsilon_n^*(\hat{\mu}, X) = \min_m \left\{ \frac{m}{n}; \sup_{\tilde{X}} \left\| \hat{\mu}(\tilde{X}) - \hat{\mu}(X) \right\| = \infty \right\} \qquad (1.3)$$

where $\tilde{X}$ is a collection of observations corrupted by replacing observations with arbitrary values. From (1.3), it can be seen that the breakdown point for a location estimator is the smallest fraction of a sample that can be corrupted by outliers before the distance between the true sample mean and the corrupted sample mean can become arbitrarily large.

The formal definition of the breakdown point for the covariance estimator, ,is given by :

$$\varepsilon_n^*(\hat{\Sigma}, X) = \min_m \left\{ \frac{m}{n}; \sup_{\tilde{X}} D\left(\hat{\Sigma}(\tilde{X}) - \hat{\Sigma}(X) \right) = \infty \right\} \qquad (1.4)$$

D(A, B) = max|/(A) – 1/(B)|, |p(A)-1 – p(B)-1|, and i(A) is A's ith ordered eigen value. In other words, the breakdown point for a covariance estimator is the smallest fraction of a sample that can be corrupted by outliers before the difference between the largest eigen values of the true and corrupted covariance estimates becomes arbitrarily large, or the difference between the smallest eigen values of the two estimates is arbitrarily close to zero. As stated by Rousseeuw and Leroy, it is preferable to employ estimators with a high breakdown point approaching the theoretical limit of 50% when estimating the mean vector and covariance matrix for a sample of data (1987). Unfortunately, the traditional mean and covariance estimators only have 1/N breakdown points, where N is the sample size (Donoho and Huber, 1983). As a result, with as little as one contaminated observation in the sample, the classical mean and covariance estimators may possibly yield unbounded estimates in the sense of (1.3) and (1.4).

Robust Distance-based Methods

There are numerous robust distance-based outlier detection methods evolved over the last two decades and the following are the findings presented in order.

a. M-Estimation Method

One of the earliest robust distance - based methods was proposed by Campbell (1980), who suggested using M-estimators to obtain robust mean vector and covariance matrix estimates. However, *M*-estimators were originally proposed by Maronna (1976), as an affine equivariant method for obtaining robust mean vector and covariance matrices for possible use in linear discrimination, principal component analysis, and outlier detection. The M-estimates of a location vector t, and a scatter matrix V, are defined as the solution to the following system of equations:

$$\frac{1}{n}\sum_{i=1}^{n} u_1\left[\left\{(x_i - t)^T V^{-1}(x_i - t)\right\}^{1/2}\right](x_i - t) = 0 \qquad (1.7)$$

$$\frac{1}{n}\sum_{i=1}^{n} u_2\left[\left\{(x_i - t)^T V^{-1}(x_i - t)\right\}^{1/2}\right](x_i - t)(x_i - t)^T = V\,, \qquad (1.8)$$

where u_1 and u_2 are functions of the Mahalanobis distance based on certain assumptions. In general, these functions serve as weighting functions that minimize the impact of outlying observations have on the mean and covariance estimates. Different forms of the weighting functions have been proposed in the literature. To find a solution for (1.7), iterative methods are typically employed but, there is no guarantee to attain the global optimum. As determined by Maronna (1976), a weakness of these estimators is a breakdown point of $1/(p+1)$, where p is the dimension of data, which can be problematic if operating in high-dimensional space.

b. MVE and MCD Methods

Rousseeuw (1983) introduced the Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) as techniques for estimating the position and dispersion of data as an alternative to the M-estimation approach with high breakdown point. The MVE technique looks for the ellipsoid with the smallest volume that covers at least h of the observations, where h is [n/2]+1 and n is the number of samples. To ensure consistency with a multivariate normal distribution, the mean vector estimate is the centre of the ellipsoid, and the covariance is the ellipsoid itself multiplied by a correction factor. Similarly, the MCD searches for the sub sample of h observations with the lowest determinant in the covariance matrix. The covariance estimate is the covariance of the h observations multiplied by a consistency factor, and the mean vector is the mean of the h observations. The Mahalanobis distance of all the data is then computed using the MVE or MCD estimations to identify outliers. The MVE and MCD have a high breakdown point of 50%, making them particularly helpful for severely polluted data. The combinatorial optimization issue that must be addressed to discover these estimators' precise answers is one of their drawbacks. To discover approximate answers, search heuristics are used in practise.

Hadi's Forward Search Method

Hadi (1992) highlighted many shortcomings with the MVE-based outlier identification technique provided by Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990). The user must first select how many sub-samples to utilise in the resampling method. This option is not apparent since it is dependent on the probably unknown percentage of outliers in the data. A second drawback is that the sub-sample covariance matrices are calculated using just p + 1 data, which may result in singularities or extremely incorrect estimates. Hadi's last point is that many sub-samples may have covariance determinants that are near to zero, leaving the user with the job of choose which sub-sample to select from the MVE estimate. Because the covariance patterns of these sub-samples may vary significantly, the resultant MVE estimations are likewise likely to differ. As a result, selecting the appropriate sub-sample is not straightforward.

Hadi presented an MVE-based, non-affine equivariant outlier identification approach that starts by calculating the vector of coordinate-wise medians for the original data to compensate for the constraints of the original MVE resampling method. The covariance matrix for the data is then estimated using the median vector. The robust Mahalanobis distances for the data are computed using these location and covariance estimations. The [(n+p+1)/2] observations with the lowest distances are selected and utilised to provide traditional mean vector and covariance estimations, as well as a new set of distances for all of the observations. The p +1 observations with the lowest distances are chosen from this most recent collection of distances to create the fundamental subset.

COMEDIAN APPROACH TO DETECT MULTIPLE OUTLIERS IN MISSING DATA

When univariate outlier detection techniques are applied to each variable, observations that are unusual in a single variable may be identified as outlying. When considering each variable measurement in respect to the other variables, an observation may be identified as an outlier in multivariate data. Multivariate outlier detection methods, for example, in clinical laboratory safety data, may highlight a patient whose laboratory measurements do not follow the same pattern of relationships as the majority of patients, despite the fact that their measurements are not found to be outlying when considered one at a time. The process of detecting outliers is an intriguing and essential element of data analysis, since it has the potential to influence inference.

Correlated Data

Since the *Comedian* method is not affine equivariant it is important to conduct a study using correlated data as its behavior depend upon the covariance structure of the data. Devlin *et al. (*1981) used a correlation matrix P for generating Monte Carlo data from different distributions of moderate dimension *(p* = 6). The matrix P = ((ρ_{ij})) has the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 \end{bmatrix} \text{with } \mathbf{P}_1 = \begin{bmatrix} 1 & 0.95 & 0.30 \\ 0.95 & 1 & 0.10 \\ 0.30 & 0.10 & 1 \end{bmatrix} \text{and } \mathbf{P}_2 = \begin{bmatrix} 1 & -0.499 & -0.499 \\ -0.499 & 1 & -0.499 \\ -0.499 & -0.499 & 1 \end{bmatrix}.$$

The matrix P has several desirable features. First, its dimension is large enough to study multivariate estimators. Secondly, the range of correlation values is large, so that differences in the abilities of the methods to detect outliers from highly correlated datasets can be checked (Devlin *et al.*, 1981).

Dataset of 100 observations were generated from an asymmetric contaminated normal which is a mixture of 100(1-a) observations from N(0,P) and 100a observations from N(5u,P), where *u* = (1, ..., 1)". The success rates of the four methods for different percentage of contamination in correlated data are observed to be same. Hence, it is clear that the *Comedian* method is as efficient as other affine equivariant methods to detect true outliers from correlated datasets. Also, Table 3.3 shows that the *Comedian* method has low false detection rates even for correlated data.

COMEDIAN APPROACH TO DETECT MIXED-TYPE OUTLIERS

Data analysis usually deals with a large number of variables being recorded and subjected to statistical scrutiny and analysis for drawing meaningful inference. The chapters covered presume that data is continuous in nature and the detection of outliers is dealt as a first step towards coherent analysis. Various methods for detecting outliers from large dimensional data sets have been studied (Barnett and Lewis, 1994) and many of these techniques assume that all attributes in the dataset are either continuous or categorical. However, many real and practical datasets are of *mixed-type* with a heterogeneous mixture of categorical (nominal) and continuous type attributes. For example, a data point representing a network flow may contain continuous type (number of bytes transferred between two hosts, length of the connection in seconds, etc.) and categorical type (service accessed, network protocol used, etc.) attributes. The problem of analysis of mixed-type datasets is not quite straightforward and needs careful consideration.

Having different attribute types in a data set makes it difficult to find relations between two attributes (for example, correlation between the attributes) and to define distance or similarity metrics for such data points. Usually, for processing datasets with a mixture of attribute types, many techniques homogenize the attributes by converting continuous attributes into categorical attributes by discretization (quantization), or converting categorical attributes into continuous attributes by applying some (arbitrary) ordering, which can lead to a loss in information and an increase in noise (Otey *et al.*, 2006). A better outlier detection system would be needed to develop meaningful and useful distance metrics in mixed-type attribute spaces and measure the dependencies between attributes of different types.

Detection of Outliers in Mixed-type Data

Inlying score of an observation is defined as the number of observations in its δ – neighbourhood ***Nhd***(δ, d) with radius δ and a distance measure d. Distance measure computes the distance of each observation from other observations and those observations that have a distance less than the radius will be considered as neighbours. The rationale behind inlying score is that, the uncontaminated observations possess large neighbourhoods than a small group of outliers. As a result, the normal observations hold a large inlying score. But, it may not be the same if the distance measure is not robust or the radius values is not properly selected. Hence it is essential to have a robust distance measure and a suitable radius value. The proposed method derives distance measure and radius value as follows.

Let ***X = [X': X"]*** be an n x p data matrix contains p_1 continuous and p_2 nominal attributes with rows $x_i^T = [x_i^I : x_i^{II}]$ $i = 1, ..., n$ and $p = p_1+p_2$. Here $\mathbf{x^I}$ is a matrix of order $n \times p_1$ with continuous attributes and X^{II} is a matrix of order $n * p_2$ with nominal attributes. Han and Kamber (2002) presented a suitable distance measure for mixed-type data which is a combination of different distance measures. Let x_i and x_j be two observations in **X**. The combined distance measure between x_i *and* x_j is given by

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{d_C\ p_1 + d_N p_2}{p_1 + p_2}, \quad i, j = 1, \dots, n \tag{3.1}$$

where d_c and d_N are the corresponding distance measures of continuous and nominal attributes.

An appropriate metric that measures the distance between two continuous observations is Mahalanobis distance. Let x_i^I and x_j^I be the two observations in $\mathbf{X^I}$. The Mahalanobis distance between x_i^I and x_j^I is defined as

$$md(\mathbf{x}_i^I, \mathbf{x}_j^I) = (\mathbf{x}_i^I, \mathbf{x}_j^I)^T\, \mathbf{C}^{-1}\, (\mathbf{x}_i^I, \mathbf{x}_j^I),\ i, j = 1, \dots, n \tag{3.2}$$

where **C** is the sample covariance matrix of $\boldsymbol{X^I}$. Small value *of md* may indicate that the corresponding observations are close to each other and are sharing a common neighborhood. However, the Mahalanobis distance suffers from the twin problems of *masking* and sw*amping.* In masking, observations holding small value for *md* may not be in same neighbourhood and in masking, neighboring observations need not possess small value fo*r md.* These problems occur due to the fact that, the sample covariance matrix S is not robust to the presence of outliers. Various estimators are available in literature and the current approach suggests the use of *Comedian* estimate proposed by Sajesh and Srinivasan (2011a). These estimates are highly robust and possess high breakdown value. Using the *Comedian* estimate S of scatter, a robust Mahalanobis distance is defined as follows

$$rd(\mathbf{x}_i^{\,I}, \mathbf{x}_j^{\,I}) = rd_{ij} = (\mathbf{x}_i^{\,I}, \mathbf{x}_j^{\,I})^T\, \mathbf{S}^{-1}\, (\mathbf{x}_i^{\,I}, \mathbf{x}_j^{\,I}),\ i, j = 1, \ldots, n. \qquad (3.3)$$

Similarly, a matching coefficient would be a suitable distance measure for nominal observations. Let x_i^I and x_j^I be the two observations in X^{II} with m matching attributes. Then the distance between x_i^I and x_j^I is defined as ,

$$mc\ (\mathbf{x}_i^{\,II}, \mathbf{x}_j^{\,II}) = (p_2 - m)/p_2,\ i, j = 1, \ldots, n. \qquad (3.4)$$

CONCLUSION

The techniques for identification of outliers and understanding its impact on data are extremely important in data analysis. It is pertinent to note that detection of outliers is at times important by itself in data analysis, as exemplified in the introductory Chapter, and further could provide distorting results.

An extensive literature is available on detection of outliers especially on univariate data (Barnett and Lewis, 1994). However, less attention has been paid in dealing with outliers present in multivariate data. The detection of multiple outliers in multivariate data is more complex than univariate data and the problem gets substantially increased with the dimension of the data. The graphical representation of high dimensional data may not be really useful to study outliers. However, multivariate data with dimension greater than two, scatter plots of all possible pair of variables could be considered but detection of outliers is not guaranteed and even if detected could be misleading. For example, an outlier containing mild but systematic errors in all of its components will remain hidden unless a suitable linear transformation of the data is performed . In addition, the number of scatter plots required gets increased enormously with the dimension of the data. Principal Component Analysis (PCA) is a popular statistical method, used to explain the inherent covariance structure of data based on a multiplicative model. The components obtained through an orthogonal transformation are linear combinations of the original variables, and often allow for a better understanding and interpretation of different sources of variation. The principal component biplot is a graphical tool to simultaneously visualize the scores and loadings of principal components obtained from the classical principal component analysis. However, PCA biplot is highly influenced by outliers and hence used limited in practice.

To overcome this difficulty and to visualize the multidimensional data into a two dimensional plane, a robust principal component biplot, called ROBPCA biplot (Sajesh and Srinivasan, 2009) has been proposed. The ROBPCA biplot making use of a robust principal component method proposed by Hubert *et al. (*2005). ROBPCA biplot can be used to analyze the correlated structure of the data. Usually, the first two components of the ROBPCA biplot explain a good amount of variability and could well be a true representation of the data. Then a well defined ellipse is superimposed on ROBPCA biplot and observations outside the ellipse are identified as outliers. The numerical study reveals that the method could very well detect outliers for a moderate sized data set with correlated structure.

Lastly, the performance of *Comedian* method gets enhanced in terms of success rate and false detection rate with the increase in the dimension of the data. The difficulty in detecting outliers from multivariate data gets compounded with the multi-cluster data as most of the available methods in literature fail to detect outliers from such type of data. Sajesh and Srinivasan (2011b) proposed a computationally efficient method to detect multidimensional outliers from multi-cluster data. The method starts with the *Subtractive clustering*, method in which the number of clusters is not known in advance. In the first phase of outlier detection, the *Comedian* method is applied to individual clusters and detect outliers from each cluster. In the second phase, the detected outliers are again checked across all the clusters. To examine the performance of *Comedian* clustering method, an extensive simulation was considered under varying number of clusters, sizes and dimension and the results revealed the efficiency of the method to detect outliers from multidimensional multi-cluster data.

In addition to the size and dimension of the multivariate data, nature of the data type of data adds additional complexity to the analysis. The multivariate data could betotally continuous, totally discrete or could be a mixed type of both continuous and discrete in nature. When the dataset contains discrete type variables most of the outlier detection methods fail to detect the exact outliers or they may not be applicable to such datasets. Sajesh and Srinivasan (2010) introduced a method to detect multidimensional outliers from mixed-type data. The proposed procedure looks into the neighbourhood of each observation and assigns an *inlying score* to the observation using a robust distance measure and a well defined radius. The distance measure used for the proposed procedure is an appropriate combination of different distance measures corresponding to different nature of variables. The efficiency of the method has been studied based on the success rate and the false detection rate using a simulation study. In addition, the method was applied to well-known real datasets to evaluate its performance.

REFERENCES

Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the International Conference on Very Large Database*s, 487–499.

Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *Proceedings of the ACM SIGMOD International Conference on Management of Data, 27(2)*, 94 – 105.

Aitchison, J. and Greenacre, M. (2002). Biplots for Compositional Data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 51(4), 375–392.

Al-Zoubi, M.B. (2009). An Effective Clustering-Based Approach for Outlier Detection. *European Journal of Scientific Research,* 28(2), 310-316.

Ammeraal, L. (1992*). Programming Principles in Computer Graphics.* Wiley, New York, USA.

Anitha, K. (2004). *Statistical Methods for Breeding Experiments Using Pedigree Information.* Ph.D Thesis, University of Madras, Chennai, India.

Atkinson, A.C. (1993). Stalactite Plots and Robust Estimation for the Detection of Multivariate Outliers. In *New Directions in Statistical Data Analysis and Robustne*ss, Morgenthaler, S., Ronchetti, E. and Stahel, W.A., Eds, Basel: Birkhauser, 1-8.

Atkinson, A.C. (1994). Fast Very Robust Methods for the Detection of Multiple Outliers. *Journal of the American Statistical Association, 8*9, 1329-1339.

Atkinson, A.C. (1985*). Plots, Transformations and Regression.* Clarendon Press, U.K.

Bay, S.D. and Schwabacher, M. (2003). Mining Distance-based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule. *Proceedings of oth annual ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Becker, C. and Gather, U. (1999). The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association*, 94, 947-955.

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

Billor, N., Hadi, A.S. and Velleman, P.F. (2000). BACON: Blocked Adaptive Computationally Efficient Outlier Nominators. *Computational Statistics and Data Analysis*, 34, 279-298.

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

Butler, R.W., Davies, P.L. and Jhun, M. (1993). Asymptotics for the Minimum Covariance Determinant Estimator. *The Annals of Statistics*, 21, 1385-1400.

Campbell, N.A. (1980). Robust Procedures in Multivariate Analysis 1: Robust Covariance Estimation. *Applied Statistics, 2*9, 231-237.

Campbell, N.A. (1989). Robust Bushfire Mapping Using NOAA AVHRR Data. *Technical Report, C*SIRO.

Caroni, C. and Prescott, P. (1992). Sequential Application of Wilks's Multivariate Outlier Test. *Applied Statistics*, 41, 355-364.

Cerioli, A. (2010). Multivariate Outlier Detection with High-Breakdown Estimators. *Journal of the American Statistical Association*, 105(489), 147-156.

Chiang, L.H., Pell, R.J. and Seasholtz, M.B. (2003). Exploring Process Data with the Use of Robust Outlier Detection Algorithms. *Journal of Process Control*, 13, 437- 449.

Cui, H., He, X. and Ng, K.W. (2003). Asymptotic Distributions of Principal Components Based on Robust Dispersions. *Biometrika*, 90, 953 -966.

Daigle, G. and Rivest, L.P. (1992). A Robust Biplot. *The Canadian Journal of Statistics*. 20, 241-255.

Daudin, J.J., Duby, C. and Trecourt, P. (1988). Stability of Principal Component Analysis Studied by the Bootstrap Method. *Statistics*, 19, 214-258.

David, H.A. (1981). *Order Statistics*. Wiley, New York.

Davies, L. (1992). The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator. *The Annals of Statistics*, 20,1828-1843.

Donoho, D.L. (1982). *Breakdown Properties of Multivariate Location Estimators*, Ph.D Qualifying Paper, Department of Statistics, Harvard University, Cambridge, MA.

Donoho, D.L. and Huber, P.J. (1983). The Notion of Breakdown Point, in *A Festschrift for Erich L. Lehmann*, Bickel, P.J., Doksum, K.A. and Hodges, J.L. Eds, Belmont, CA, 157-184.

Eaton, M.L. (1983). Isotropic Distributions. In *Encyclopedia of Statistical Sciences,* Kotz, S., Johnson, N.L. and Read, C.B., Eds, Wiley, New York, 265-267.

Fang, K.T. and Wang, Y. (1994). *Number Theoretic Methods in Statistics*. Chapman and Hall, London.

Fox, A.J. (1972). Outliers in Time Series. *Journal of the Royal Statistical Society, Series B,* 34, 340-363.