An Overview of Multiple Outliers in Multidimensional Data

T. A. Sajesh[1] and M.R. Srinivasan[2]

[1]Department of Statistics, St. Thomas' College, Thrissur, India
[2]Department of Statistics, University of Madras, Chennai, India
Corresponding Author: sajesh.abraham@yahoo.com

Received: 05 December 2012 / Revised: 27 September 2013 / Published: 15 November 2013

ABSTRACT

The process of detection of outliers is an interesting and important aspect in the analysis of data, as it could impact the inference. Literature is abundant with procedures for detection and testing of single outliers in sample data. However, the presence of two or more outliers in multivariate data would render the detection and testing process more complicated as majority of outliers are invisible to many of the methods. This is due to the masking effect, and regular classical and related methods being found unsuitable for use of outlier identification techniques. The difficulty of detection increases with the number of outliers and the dimension of the data because the outliers can be extreme in any growing number of directions. An overview of multivariate outlier detection methods are provided in this study because of its growing importance in a wide variety of practical situations.

Keywords: Outlier detection, Breakdown value, Mahalanobis distance, Robust statistics.

1. Introduction

Statisticians have always been interested in finding "outlying", "unusual", or "unrepresentative" observations for many years as a precursor to data analysis. Data incorrectly entered or that do not belong to the population from which the rest of the data came can bias the estimates and give misleading results. Methods have been devised to identify and/or accommodate outlying observations in a variety of situations. With recent advances in technology, scientists are collecting large data sets, and the analyst is getting deeper to unravel the mysteries of data. So, it is

important to have a good methodology for dealing with rogue observations that might not be noticed in a typical data analysis.

The basic definition of an outlying observation is a data point or points that do not fit the model of the rest of the data. Specific definitions are given such as:

An outlier is a point such that "in observing a set of observations in some practical situation one (or more) of the observations 'jars' stands out in contrast to other observations, as extreme value." [1].

An outlying observation, or 'outlier', is one that appears to deviate markedly from other members of the sample in which it occurs.[2].

An outlier is "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism." [3].

However, the words "stands out", "appears to deviate", and "arouse suspicions" imply some kind of subjectivity or preconceived ideas about what the data should look like. Though formal methods also often rely on distributional assumptions, formal methods will cut down on the amount of subjectivity used in data analyses that employ outlier detection methods.

There are two basic reasons to search for outliers: i) the interest in the outliers for their own sake, and ii) the outliers could influence the results from the rest of the data. In 1949 in England, the case of Hadlum vs. Hadlum provides a good example of interest in outliers for their own sake. Mr. Hadlum appealed the rejection of an earlier petition for divorce on grounds of adultery. Mrs. Hadlum had given birth to a child (who she claimed was fathered by her husband) on August 12, 1945; 349 days after Mr. Hadlum had left the country. The average gestation period for a human female is 280 days, and so the question arose regarding 349 days being simply as a large observation or does that data point belong to another population, namely one of women who conceived much later than August 28, 1944 [1]. Another example where outliers themselves are of primary importance involves air safety, as discussed in [4]. Further applications of outlier identification to homeland security are described in an article by Banks [5].

Conversely, imagine a scientist studying a certain type of mosquito. If there were other types of mosquitoes in his data collection, he would not be interested in their characteristics, he would simply want to remove the observations or ensure that the observations do not influence the statistical estimates of the original population. In such a situation, the techniques should accommodate the outliers but need not detect and reject them in the estimation and are hence called robust. Thus, robustness signifies insensitivity to small deviations from the assumptions [6].

Along with identifying or accommodating outliers, some idea of why or how the outliers arose is important. Barnett and Lewis [1] classify the types of variation into three groups.

i. *Inherent Variability* – is the natural variability in any data set.
ii. *Measurement Error* – includes limitation of the measuring device as well as any recording error done by the scientist.
iii. *Execution Error* – include situations with observations which are not in the population of interest or situations when a biased or misjudged sample is used.

If the variability is due to measurement error or execution error, the point should probably be identified from the sample. However, if the variability is due to inherent variation, the point should remain.

2. Univariate Outliers

In univariate data, the concept of outlier seems relatively simple to define. Outliers are those points located "far away" from the majority of the data and "probably do not" follow the assumed model. A simple plot of the data, such as scatter plot, stem-and-leaf plot, *QQ*-plot, etc., can often reveal which points are outliers. This is sometimes called the "interoccular test" because it hits between the eyes.

Tukey [7] introduced the most popular graphical procedure called a boxplot for detecting outliers from univariate data. The boxplot rule declares observations as outliers if they lie outside the interva

$$(Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)) \quad (2.0.1)$$

where Q_i is the i^{th} quartile. The common choices for k is 1.5 for flagging "out" values and 3.0 for flagging "far out" observations. Since this rule is not sample-size dependent, the probability of declaring outliers when none exist changes with the number of observations. In that sense, it differs from standard outlier identification rules, which are set at α probability of identifying outliers when none exist.

Hoaglin *et al*. [8], showed that the popular boxplot outlier labeling rule is highly liberal with a 50% chance of labeling at least one outlier for data coming from a random normal sample of size 75. Hoaglin and Iglewicz [9], modified the rule to make it sample-size dependent, so that this probability stays at 5% for normal samples up to 300 observations. Banerjee and Iglewicz [10], expanded this modified rule to handle large sample cases and also a great variety of continuous univariate distributions. Kimber [11], slightly modified the standard boxplot outlier-labeling rule for skewed distributions by replacing;

$$Q_3 + k(Q_3 - Q_1) \text{ by } Q_3 + k(Q_3 - M) \text{ and } Q_1 - k(Q_3 - Q_1) \text{ by } Q_1 - k(M - Q_1) \quad (2.0.2)$$

where M is the sample median. Kimber also used $k = 1.5$ and studied the exponential distribution, including cases where the data is right censored, using the Kaplan–Meier estimator to estimate the median and quartiles for censored data. Van der Loo [12] introduced two univariate outlier detection methods. In both methods, the distribution of the bulk of observed data is approximated by regression of the observed values on their estimated *QQ*-plot positions using a model cumulative distribution function.

The study of outliers in structured situations like regression models and designed experiments has been carried out by numerous authors including Gentleman and Wilk ([13], [14]), John and Draper [15], Prescott [16], and John [17] and are based on residuals. Balasooriya *et al.* [18] carried out an empirical study to identify the best of seven commonly used methods for identifying outliers in linear regression models based on several data sets. The methods they compared are due to Tietjen *et al.* [19], Prescott [16], Andrews and Pregibon [20], Cook and Weisberg [21], Cook [22], and Draper and John [23]. On the basis of their study, they observed that the methods do not always agree and suggested a judicious *combination of procedures*. Their empirical studies also revealed that the results tend to strongly disagree in the case of multiple outliers. Balasooriya and Tse [24] studied the relative performance of five widely used test statistics for detecting outliers using Monte Carlo method. Through this study, they identified that the test statistic based on studentized residuals proposed by Tietjen *et al.* [19] is the best procedure for detecting a *single outlier*.

3. Multivariate Outliers

Multivariate outliers pose bigger challenges than univariate data as simple visual detection of multivariate outliers is virtually impossible because the outliers do not "stick out" on the end [25]. Even plotting the data in bivariate form with a systematic rotation of coordinate pairs will not help. Barnett and Lewis [1] and Beckman and Cook [26], presented several key concepts that point to the relevance of multivariate outlier detection methods for anomaly detection.

Breakdown point is an important measure that is used to describe the resistance of robust estimators in the presence of outliers. Following Hodges [27] and Hampel ([28], [29]), breakdown point of an estimator is the fraction of arbitrary contaminating observations that can be presented in a sample before the value of the estimator can become arbitrarily large. Lopuhaä and Rousseeuw [30], have presented more formal definitions of the breakdown point for location and

covariance estimators. For a location estimator, $\widehat{\boldsymbol{\mu}}$, at a collection of observations $\boldsymbol{X}$, the breakdown point $\varepsilon_n^*(\widehat{\boldsymbol{\mu}}, \boldsymbol{X})$ is defined as:

$$\varepsilon_n^*(\widehat{\boldsymbol{\mu}}, \boldsymbol{X}) = \min_m \left\{ \frac{m}{n}; \sup_{\widetilde{\boldsymbol{X}}} \left\| \widehat{\boldsymbol{\mu}}(\widetilde{\boldsymbol{X}}) - \widehat{\boldsymbol{\mu}}(\boldsymbol{X}) \right\| = \infty \right\} \tag{3.1}$$

where $\widetilde{\boldsymbol{X}}$ is a collection of observations corrupted by replacing observations with arbitrary values. From (3.1), 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, $\widehat{\boldsymbol{\Sigma}}$, is given by :

$$\varepsilon_n^*(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{X}) = \min_m \left\{ \frac{m}{n}; \sup_{\widetilde{\boldsymbol{X}}} D\left(\widehat{\boldsymbol{\Sigma}}(\widetilde{\boldsymbol{X}}) - \widehat{\boldsymbol{\Sigma}}(\boldsymbol{X}) \right) = \infty \right\} \tag{3.2}$$

where $D(\mathbf{A}, \mathbf{B}) = max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}|\}$, and $\lambda_i(\mathbf{A})$ is the i^{th} ordered eigen value of $\mathbf{A}$. In other words, (3.2) states that 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 covariance estimate and that of the corrupted covariance estimate becomes arbitrarily large, or the difference between the smallest eigen values of the two estimates is arbitrarily close to zero. In the context of estimating the mean vector and covariance matrix for a sample of data, it is advantageous to use estimators with a high breakdown point touching the theoretical limit of 50%, as explained by Rousseeuw and Leroy [31]. Unfortunately, the breakdown points for the classical mean and covariance estimators are only 1/*N*, where *N* is the sample size [32]. Hence, the classical mean and covariance estimators can potentially produce unbounded estimates, in the sense of (3.1) and (3.2), with as little as one contaminating observation present in the sample.

The *influence function* is also an important robust measure, which measures the effect on an estimator of adding a small mass at a specific point [33]. Robust estimators ideally have a bounded influence function, which means that a small contamination at any point can only have a small effect on the estimator [34]. As discussed in Hampel *et al.* [33], the importance of the influence function lies in the fact that it can describe the effect of an infinitesimal contamination at the point $\boldsymbol{x}$ on the estimate *T*, standardized by the mass of the contaminant. It gives us a picture of the asymptotic bias caused by contamination in the data.

Another desirable property of a robust estimator is affine equivariance. A location estimator $\widehat{\boldsymbol{\mu}} \in \boldsymbol{R}^p$ is affine equivariant if and only if for any vector $\boldsymbol{b} \in \boldsymbol{R}^p$ and any non-singular $p \times p$ matrix $\boldsymbol{A}$,

$$\widehat{\boldsymbol{\mu}}(\boldsymbol{AX} + \boldsymbol{b}) = \boldsymbol{A}\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) + \boldsymbol{b} \tag{3.3}$$

A scale estimator $\widehat{\boldsymbol{\Sigma}} \in PDS(p)$ (the set of positive-definite symmetric $p \times p$ matrices) is affine equivariant if and only if for any vector $\boldsymbol{b} \in \boldsymbol{R}^p$ and any nonsingular $p \times p$ matrix $\boldsymbol{A}$,

$$\widehat{\boldsymbol{\Sigma}}(\boldsymbol{AX} + \boldsymbol{b}) = \boldsymbol{A}\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X})\boldsymbol{A}^{\boldsymbol{T}} \tag{3.4}$$

If an estimator is affine equivariant, stretching or rotating the data will not affect the estimator. Dropping this requirement greatly increases the number of available estimators, and in many cases, non-affine equivariant estimators have superior performance to affine equivariant estimators.

In addition to estimator breakdown, the phenomenon of outlier masking also argues for the use of outlier resistant detection methods for detecting multidimensional outliers. Masking refers to the condition of very strong outliers distorting non-robust mean and covariance estimates to such a degree that weaker outliers appear ordinary in terms of their Mahalanobis distances. If there is one or more distant outlier and one or more not so distant outlier in the same direction, the more distant outlier(s) could significantly shift the mean in that direction, and also increase the standard deviation, to such an extent that the lesser outlier(s) falls less than 2 or 3 standard deviations from the sample mean, and goes undetected. The degree of masking is measured in terms of an increase in Type II error, or false negatives, since observations that are truly outlying are classified as part of the uncontaminated population of data.

Becker and Gather [35], developed the masking breakdown point of outlier detection method that specifies the smallest fraction of outliers in a sample that can induce the masking affect. Becker and Gather prove that the masking breakdown point for an outlier detection method that uses a mean and covariance estimator is bounded by the breakdown points of these two estimators. Further, if the two estimators have the same breakdown point, then the masking breakdown point of the detector is equal to the estimator breakdown point. An immediate conclusion that can be drawn from these findings is that non-robust Mahalanobis distance-based outlier detection methods can be affected by masking in the presence of a single outlying observation.

Further reason for employing multivariate outlier detection methods for anomaly detection is to combat the swamping effect. Masking refers to the increase of Type II error due to the presence of outliers and swamping refers to the increase in Type I

error caused by outliers. Hadi [36], observed that not all observations with large [Mahalanobis distance] values are necessarily outliers. For example, a small cluster of outliers will attract [the mean vector] and will inflate [the covariance estimate] in its direction and away from some other observations which belong to the pattern suggested by the majority of observations. To ensure against this source of false alarms, multivariate outliers detection methods should be employed that use robust estimation methods for the mean vector and covariance matrix. Following this strategy helps ensure that the false alarm rate for an anomaly detection method is inline with the accepted α-level for the method.

Various methods have been proposed over the years to detect outliers and are broadly classified into: *robust distance-based methods*, and *non-traditional methods*. The robust distance methods use some form of robust estimation to obtain mean vector and covariance estimates for the data. The Mahalanobis distance is then computed for each observation using these robust estimates, and observations whose distances exceed a critical value – generally from the Chi-square distribution if the data is multivariate normal – are labeled as outliers. For the non-traditional methods, some alternative statistic is exploited that is presumably better at revealing outliers or computationally easier than distances based on robust mean and covariance estimates. Both the methods are discussed in detail in the following sections.

3.1. 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.

3.1.1. M-Estimation Method

One of the earliest robust distance-based methods was proposed by Campbell [37], who suggested using *M*-estimators to obtain robust mean vector and covariance matrix estimates. However, *M*-estimators were originally proposed by Maronna [38], 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 $\boldsymbol{t}$, and a scatter matrix $\boldsymbol{V}$, are defined as the solution to the following system of equations:

$$\frac{1}{n}\sum_{i=1}^{n} u_1\left[\{(\boldsymbol{x}_i-\boldsymbol{t})^T\boldsymbol{V}^{-1}(\boldsymbol{x}_i-\boldsymbol{t})\}^{\frac{1}{2}}\right](\boldsymbol{x}_i-\boldsymbol{t})=0 \tag{3.5}$$

$$\frac{1}{n}\sum_{i=1}^{n} u_2\left[\{(\boldsymbol{x}_i-\boldsymbol{t})^T\boldsymbol{V}^{-1}(\boldsymbol{x}_i-\boldsymbol{t})\}^{\frac{1}{2}}\right](\boldsymbol{x}_i-\boldsymbol{t})(\boldsymbol{x}_i-\boldsymbol{t})^T=\boldsymbol{V} \tag{3.6}$$

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 (3.5), iterative methods are typically employed but, there is no guarantee to attain the global optimum. As determined by Maronna [38], 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.

3.1.2. MVE and MCD Methods

As an alternative to the *M*-estimation method with high breakdown point, Rousseeuw [39], proposed the Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) as methods for estimating the location and scatter of the data. The MVE method searches for the minimal volume ellipsoid that encompasses at least h of the observations, with h taken as $[n/2] + 1$, where n is the number of samples. The mean vector estimate is the center of the ellipsoid, and the covariance is the ellipsoid itself multiplied by a correction factor to achieve consistency with a multivariate normal distribution. In a similar manner, the MCD looks for the sub-sample of h observations whose covariance matrix has the smallest determinant. The mean vector is then taken as the mean of the h observations, and the covariance estimate is the covariance of the h observations multiplied by a consistency factor. The MVE or MCD estimates are then used to compute the Mahalanobis distance of all the observations to detect outliers. The advantage of the MVE and MCD is the high breakdown point of 50%, and hence very useful for highly contaminated data. A disadvantage of these estimators is the combinatorial optimization problem that must be solved to find their exact solutions. In practice, search heuristics are employed to find approximate solutions.

A practical means for searching an approximate MVE solution was proposed by Rousseeuw and Leroy [31], and again by Rousseeuw and van Zomeren [40]. This method – referred to as the resampling method – entails drawing m sub-samples of size $p + 1$ from the original data, where m is chosen to ensure a high probability that at least one sub-sample will be free of outliers. For each sub-sample, the covariance matrix is computed and either inflated or deflated to include h of the observations from the original sample. The volumes of each of the m resulting ellipsoids are then approximated, and the one with the minimum volume is used to form the MVE estimate.

To improve the efficiency of the MVE estimate, Rousseeuw and Leroy [31] go on to recommend a *reweighting* step in which the mean vector and covariance matrix are recomputed using only the observations whose Mahalanobis squared distance

relative to the MVE mean vector and covariance matrix fall below a suitable quantile of a Chi-square distribution with *p* degrees of freedom. This reweighting step is also recommended by Rousseeuw and van Zomeren [40], while Lopuhaä and Rousseeuw [30], show that it preserves the breakdown point of the MVE.

When the MVE and MCD estimation methods were originally proposed by Rousseeuw [39], the MVE received initial attention for outlier detection because it was computationally less expensive to find an approximate MVE solution. However, Butler *et al*. [41], showed that the MCD has better statistical efficiency than the MVE since the MCD is asymptotically normal. Additionally, Davies [42], showed that the MVE has a lower convergence rate than the MCD. According to Rousseeuw and Van Driessen [43], theoretical findings combined with the need for accurate estimators for use in outlier detection schemes, the MCD began to gain favor over the MVE as the preferred robust estimator for outlier detection. The main drawback to using the MCD, however, was the high computational complexity involved with searching the space of half-samples of a dataset to find the covariance matrix with minimum determinant.

To address this problem, Rousseeuw and Van Driessen [43], proposed the FAST-MCD outlier detection method that uses a key theoretical finding in conjunction with a partitioning method to rapidly search for an approximate MCD solution. The primary theorem proved by Rousseeuw and Van Driessen states that if one starts with a half-sample of data, orders the entire data set based on Mahalanobis distances derived from the half-sample's mean vector and covariance matrix, and selects a new half-sample from the observations with smallest distances, the covariance determinant of the new half-sample will be less than or equal to the old half-sample covariance determinant. By repeatedly applying this theorem to a dataset – a process referred to as a C-step – it is possible to converge to at least a local optimal MCD solution. A further finding based on experimental results indicates that if the starting half-sample is capable of converging to a good solution, the covariance determinant will begin to rapidly converge after only two C-steps.

3.1.3. Stahel - Donoho Estimator

In addition to suggesting the MVE and MCD estimators for use in robust distance outlier detectors, Rousseeuw and Leroy [31] also allude to using Stahel-Donoho estimators in the robust distance computation. These estimators, proposed independently by Stahel [44] and Donoho [45], compute the mean vector and covariance matrix by assigning decreasing weight to observations that are outlying relative to some projection of the data to univariate space. Specifically, *outlyingness* of an observation $\boldsymbol{x}_i$ is defined to be:

$$u_i = \sup_{\|\boldsymbol{v}\|} \frac{\left|\boldsymbol{v}^T\boldsymbol{x}_i - \underset{j}{\text{med}}(\boldsymbol{v}^T\boldsymbol{x}_j)\right|}{\underset{k}{\text{med}}\left|\boldsymbol{v}^T\boldsymbol{x}_k - \underset{j}{\text{med}}(\boldsymbol{v}^T\boldsymbol{x}_j)\right|} \tag{3.7}$$

where $\boldsymbol{v}$ is a p-dimensional projection vector. Upon determining the u_i for all observations, the mean vector and covariance matrix are estimated as:

$$T(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} w(u_i)\boldsymbol{x}_i}{\sum_{i=1}^{n} w(u_i)} \tag{3.8}$$

$$V(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} w(u_i)\big(\boldsymbol{x}_i - T(\boldsymbol{X})\big)\big(\boldsymbol{x}_i - T(\boldsymbol{X})\big)^T}{\sum_{i=1}^{n} w(u_i)} \tag{3.9}$$

where $w(u_i)$ is a positive, decreasing weighting function.

The Stahel-Donoho estimator is an attractive robust estimator because it has a high breakdown point which asymptotically approaches 50%, as shown by Donoho [45]. However, as explained by Rousseeuw and Leroy [31], the primary difficulty with these estimators is the computation of the outlyingness values. Apparently, no satisfactory method has been proposed to find these values, thereby preventing these estimators from experiencing any practical use for outlier detection. However, Gasko and Donoho [46] propose a method that uses these estimators to identify leverage points in multiple regression data.

3.1.4. Hadi's Forward Search Method

Returning to the MVE-based outlier detection method proposed by Rousseeuw and Leroy [31] and Rousseeuw and van Zomeren [40], Hadi [36] identified several limitations with the approach. First, the user must decide upon the number of sub-samples to use in the resampling scheme. This choice is not obvious since it depends on the presumably unknown fraction of outliers that exist in the data. A second limitation is that the covariance matrices for the sub-samples are estimated using only p + 1 observations which could lead to singularities or highly inaccurate estimates. The final problem highlighted by Hadi is that several of the sub-samples may have covariance determinants close to zero, leaving the user with the task of choosing which sub-sample to use to from the MVE estimate. Since these sub-samples may have considerably different covariance structures, the resulting MVE estimates are also likely to be different. Thus, choosing the correct sub-sample is not obvious.

To correct for the limitations of the original MVE resampling method, Hadi proposed an MVE-based, non-affine equivariant outlier detection method that begins by computing the vector of coordinate-wise medians for the original data. The median vector is then used to estimate the covariance matrix for the data. These

location and covariance estimates are then used to compute robust Mahalanobis distances for the observations. The [(n+p+1)/2] observations with the smallest distances are identified and used to form classical mean vector and covariance estimates and a new set of distances for all the observations. From this latest set of distances, the p +1 observations with the smallest distances are selected to form what is referred to as the basic subset. This basic subset is analogous to a sub-sample in the MVE resampling method with two notable differences. First, the basic subset is composed of observations closest to the centroid of the sample as determined by the robust, coordinate-wise median Mahalanobis distances. Second, there is only one basic subset in Hadi's method as opposed to potentially hundreds of subsamples in the resampling MVE method. This considerable reduction in the number of subsets makes Hadi's method less computationally complex and faster to execute.

3.1.5. Atkinson's Forward Search Method

Sharing the same concerns with the MVE resampling method as Hadi, Atkinson [47] proposed an affine equivariant forward search algorithm similar in nature to Hadi's method. Atkinson's forward search method begins by randomly selecting a subset of $m = p$+1 observations and using this subset to estimate a mean vector and covariance matrix. The covariance matrix is inflated or deflated to include h of the original observations, and the volume of the resulting matrix is recorded. The adjusted covariance matrix is then used to compute the Mahalanobis squared distances for all observations and the m+1 observations with the smallest distances are used to repeat the process, while any observations whose squared distances exceed a critical Chi-Square threshold are identified as potential outliers. When $m = n$, the entire process is repeated with a new random subset of $m = p$+1 observations. After executing the algorithm through the desired number of random starting subsets, the adjusted covariance matrix that gave the smallest volume over all trials can be used for the final robust mean and covariance estimates and subsequent outlier detection. However, Atkinson does not recommend identifying outliers in this manner. Rather, he uses a graphical method known as *stalactite plots* to analyze which of the observations consistently emerged as outliers in each stage of the algorithm. Atkinson's method is well illustrated in Atkinson [48].

3.1.6. Hawkins' Feasible Solution Algorithm

Motivated by the need to use efficient starting solutions for *M*-estimation and other iterative robust estimators, Hawkins [49] proposed the Feasible Solution Algorithm (FSA) for obtaining approximations to Rousseeuw's MCD estimator. Hawkins also suggests that the MCD estimate resulting from the FSA can be used to detect outliers using the usual robust distance scheme. The FSA begins by first assuming that there are at most k outliers in the data. A random sample of (n - k) observations

is then selected from the original sample of n observations, with the remaining k observations trimmed from the data. The randomly selected observations are used to form an initial mean vector and covariance estimate along with the respective covariance determinant.

Next, for each possible pair of observations with one observation coming from the randomly selected subset and the other from the trimmed subset, an updating formula provided by Hawkins is used to determine the reduction in covariance determinant if the pair of observations is interchanged between subsets. The pair of observations that produces the greatest reduction in the covariance determinant are then swapped and the process repeated until no swaps can be identified that reduce the determinant value. The subset of $n - k$ observations that results with no scope for further improvement is referred to as a feasible solution. The entire process is then repeated to find additional feasible solutions. The final MCD estimate is obtained from the feasible solution that produced the smallest covariance determinant.

3.1.7. Hybrid Algorithm

The robust distance outlier detection methods discussed thus far follow one of three strategies: 1) use of what Rocke and Woodruff [50] refer to as smooth estimators, such as *M*-estimators or Stahel-Donoho estimators; 2) use of combinatorial estimators such as the MVE or MCD; and 3) use of forward search methods as proposed by Hadi and Atkinson. In an effort to unify these strategies under one outlier detection method, Rocke and Woodruff [50], proposed a hybrid algorithm for the detection of outliers. This method culminates the research of Rocke and Woodruff [51], Woodruff and Rocke [52], Woodruff and Rocke [53]and Rocke [54]. The high breakdown point, affine equivariant detection method is composed of two phases. The objective of Phase I is to obtain a robust estimate of the data set's location and shape. This estimate is achieved by first using Hawkins' FSA to obtain an approximate MCD estimate of the location and shape. The MCD estimate is then used for the starting point of Atkinson's forward search method as opposed to the mean vector and covariance matrix of a random subset of p+1 points originally suggested by Atkinson. The non-outlying points identified by Atkinson's method are used to compute the starting mean vector and covariance matrix estimates for a modified, high breakdown point *M*-estimation method proposed by Rocke [54]. The rationale for obtaining the final estimates in this manner is that the forward search method achieves better results given a good starting point, while *M*-estimation is also more likely to find the globally optimal solution if the initial estimate is close to this solution. An additional feature of the Phase I process is a partitioning scheme designed to counter the fact that MCD computations grow exponentially with the sample size. Rather than attempt to apply the compound MCD, forward selection,

and *M*-estimation method to the entire data set, the original data is randomly partitioned into a user-specified number of subsets. Robust estimates are then obtained for each subset and the covariance estimate with minimum determinant is used for next Phase.

Phase II of the compound estimation method involves computing the Mahalanobis squared distances for all the observations using the robust estimates from Phase I, scaling these squared distances so that they are consistent with distances obtained from multivariate normal data, and comparing the scaled distances to a suitable threshold from a Chi-square distribution with *p* degrees of freedom.

3.1.8. Smallest Half-Volume and Resampling by Half-Means Methods

Rocke and Woodruff's hybrid algorithm represents a combination of two somewhat theoretical approaches to detecting outliers. The main drawback of MCD and *M*-estimation strategy for robust distance detection is their large computational burden that limits their utility relative to large-scale problems. As a less-formal, intuitive alternative for outlier detection on large datasets, Egan and Morgan [55] propose the Smallest Half-Volume (SHV) method. The basic premise behind the SHV method is that good observations in a dataset will tend to cluster closely together in Euclidean space. To identify a cluster of good data, the method begins by mean-centering and standardizing each column of the data matrix using the respective column mean and standard deviation. This process is referred to as auto-scaling. Using the auto-scaled data, an $n \times n$ distance matrix is formed in which element d_{ij} is the Euclidean distance from observation i to observation j. Thus, each column of the distance matrix records how close observation j is to all other observations. With this idea in mind, each column of the distance matrix is sorted in ascending order. For each sorted column, the sum of the first $n/2$ distances is computed. The column with the smallest sum is identified, and the $n/2$ observations used in computing this column's sum are labeled as good data. The good data are then used to form a robust mean vector and covariance matrix, and to re-perform the auto-scaling procedure. To detect outliers, the mean vector and covariance estimates are used as robust inputs to the classic Mahalanobis distance detector.

In the same article in which the SHV method is proposed, Egan and Morgan [55], also developed the Resampling by Half-Means (RHM) method for detecting outliers. This method makes use of the auto-scaling concept to create samples of robust distances for the observations. The RHM method begins by randomly selecting $n/2$ observations from the dataset without replacement. Each of the selected observations is used to form a row of the matrix $\mathbf{X}_{(i)}$, where i denotes the iteration of the method. The mean and standard deviation are computed for each column of $\mathbf{X}_{(i)}$. These estimates are then used to auto-scale the original data matrix.

The magnitude of each row of the auto-scaled matrix is computed, which is equivalent to computing the distance of each auto-scaled observation to the centroid of the data. The distances for the n observations are saved in the vector $\mathbf{l}_{(i)}$ which, in turn, constitutes the i^{th} column of a matrix $\mathbf{L}$. This process is repeated for iteration i+1 until the desired number of iterations is achieved. After the last iteration is complete, each column of $\mathbf{L}$ is sorted in ascending order. For each of the sorted columns, the observations corresponding to the largest 5% of the distances are identified. Outliers are identified as those observations whose distances appear in the upper 5% of distances an unusually large number of times. Unfortunately, no guidance is provided as to how many appearances are indicative of an outlier and thus, the method ultimately relies on subjective judgment by the analyst.

3.1.9. Bivariate Boxplot Method

An informal method for detecting outliers in univariate data is to construct a boxplot that visually depicts the location, spread, and skewness of the data. Zani *et al*. [56], develop a method for building a bivariate boxplot and suggest how it may be used to mind multivariate outliers. To build the bivariate boxplot for pair of variables, the inner region for the plot – analogous to the univariate boxplot's inter-quartile region – is determined through the use of convex hull peeling originally proposed by Bebbington [57]. Convex hull peeling entails identifying the observations on the convex hull of the bivariate data cloud, trimming these observations from the dataset, and repeating the process until only a desired percentage of the original observations remain. For the purpose of the bivariate boxplot, Zani *et al*. suggest trimming the data until 50% of the observations remain. These observations define the inner region for the boxplot. To ensure a smooth ellipse that visually depicts this inner region, Zani *et al*. use the method of B-splines [58] to fit a curve to the convex hull of the inner region. The centroid for the boxplot is computed as the arithmetic mean of the observations contained in the inner region.

To detect multivariate outliers, Zani *et al*. recommend constructing a bivariate boxplot for every pair of variables. Any observation that is outside the 90% convex hull in any of the plots is removed from the data set. The remaining observations are then used as the starting point for the forward search method of Hadi ([36], [59]) or Atkinson [47]. The authors claim that using bivariate boxplot in this manner make the forward search more computationally efficient, presumably because the initial basic subset for the search should contain considerably more than p+1 points.

3.1.10. BACON Method

The desire to find an outlier detection method that is applicable to very large datasets is echoed by Billor *et al*. [60]. However, where the FASTMCD method attempts to use nesting and C-steps to search for an optimal solution, Billor *et al*.

make two observations concerning robust distance computation as a guide to developing the Blocked Adaptive Computationally Efficient Outlier Nominator (BACON). The first observation is that the added computational complexity of trying to find optimal robust estimators may not be justified by significantly better outlier detection. The second observation is that insisting upon a completely affine equivariant method may add substantial computational complexity to an algorithm without a proportional improvement in the detection of outliers. Using these two observations, Billor *et al*. develop BACON as a method that "abandons" optimality conditions in favor of a very fast outlier detection strategy that can be run in a non-robust, affine equivariant mode with breakdown point of 20%, or in a robust, near-affine equivariant mode with a breakdown point of 40%.

The BACON method is derived from the forward search method of Hadi ([36], [59]), and begins its search for outliers in much the same manner by selecting an initial basic subset of good observations. The manner in which the initial basic subset is chosen depends on whether the user wishes to have a lower breakdown point method that is affine equivariant, or a high breakdown point method that is not completely affine equivariant. In the former case, the initial basic subset contains the p+1 observations with the smallest Mahalanobis distances relative to the mean vector and covariance matrix for the entire dataset. In the latter case, the basic subset is formed from the p+1 observations with smallest distances relative to the component-wise median of the observations and the covariance matrix derived from this median vector. Using the component-wise median makes the BACON method more robust to outliers at the expense of affine equivariance since the median estimator is not affine equivariant. Once the initial basic subset is selected, its mean vector and covariance matrix are estimated and used to compute Mahalanobis distances for all observations. Once these distances are obtained, they can be compared to the square root of an appropriate quantile from the Chi-Squared distribution with p degrees of freedom.

3.1.11. Kurtosis Method

In spite of its computational and other difficulties, the Stahel-Donoho estimator was nevertheless important in leading to the development of other estimators. One way to reduce the extreme computational burden is to decrease the number of examined projections. Peña and Prieto [61], presented the Kurtosis method that projects the data onto a set of 2p directions, where p is the dimension of the data. These directions are chosen so as to maximize or minimize the kurtosis coefficient of the projected data. The kurtosis coefficient is a measure of how peaked or flat the distribution is. Datasets with high kurtosis tend to have a sharp density peak near the mean, decline rather rapidly, and have heavy tails. Symmetric outliers lead to heavy tails and thus, higher kurtosis. A small amount of asymmetric contamination would

also increase the kurtosis. The kurtosis coefficient is also affected by modality, as large number of asymmetric outliers would start introducing bimodality, leading to a very low value of kurtosis.

Peña and Prieto [61], thus argue that searching for outliers along the projections that maximize and minimize the kurtosis coefficient would be very promising. The exact solution of the kurtosis maximization and minimization problems requires a global solution, which is not efficient, so they settle instead for p local maximizers and p local minimizers. They show that computing a local maximizer or minimizer corresponds to the finding either (1) the direction from the center of the data straight to the outliers or (2) a direction orthogonal to it. As it not known which of these two directions have been found, the data need to be projected onto a subspace orthogonal to the computed directions, and another local solution obtained. This process has to be repeated a maximum of p times to find the desired direction, yielding a total of $2p$ examined directions where p directions for the maximum and p for the minimum. For each of these $2p$ directions, Peña and Prieto [61], determine outlyingness based on the univariate median and Median Absolute Deviation (MAD). If a point is an outlier in any of these directions, that is, considering its maximum deviation from the median, it is labeled a potential outlier. The mean and covariance are then computed based on all points not considered to be potential outliers, followed by a robust Mahalanobis distance for each point. If the Mahalanobis distance for any point exceeds the critical value of a χ^2 distribution with p degrees of freedom, it is declared an outlier.

3.1.12. OGK Method

Maronna and Zamar [62], proposed an Orthogonalized Gnanadesikan-Kettenring (OGK) estimator by a general method to obtain positive-definite and approximately affine-equivariant robust scatter matrices starting from any pair-wise robust scatter matrix. This method was applied to the robust covariance estimate of Gnanadesikan and Kettenring [25]. The resulting multivariate location and scatter estimates are called orthogonalized Gnanadesikan-Kettenring (OGK) estimates and are calculated as follows:

1. Let $m(.)$ and $s(.)$ be robust univariate estimators of location and scale
2. Construct $\boldsymbol{y}_i = \boldsymbol{D}^{-1}\boldsymbol{x}_i$ for $i = 1, \ldots, n$ with $\boldsymbol{D} = \text{diag}(s(X_1), \ldots, s(X_p))$.
3. Compute the matrix $\mathbf{U} = (u_{jk})$ with
$$u_{jk} = \begin{cases} \frac{1}{4}\left(\sigma(Y_j+Y_k)^2 - \sigma(Y_j-Y_k)^2\right) & j \neq k \\ 1 & j = k \end{cases}$$
4. Compute the matrix $\boldsymbol{E}$ of eigenvectors of $\boldsymbol{U}$ and
   a) project the data on these eigenvectors, i.e. $\boldsymbol{V}=\boldsymbol{Y}\,\boldsymbol{E}$;
   b) compute 'robust variances' of $\boldsymbol{V}= (\boldsymbol{V}_1,\ldots,\boldsymbol{V}_p)$, i.e. $\boldsymbol{\Lambda}= \text{diag}(s^2(\boldsymbol{V}_1),\ldots,s^2(\boldsymbol{V}_p))$;

c) set the p × 1 vector $\widehat{\boldsymbol{\mu}}(\boldsymbol{Y})$ =$\boldsymbol{Em}$ where $\boldsymbol{m}$ = $(m(\boldsymbol{V}_1),...,m(\boldsymbol{V}_p))^T$, and compute the positive definite matrix $\widehat{\boldsymbol{\Sigma}}$ $(\boldsymbol{Y})$ = $\boldsymbol{E\Lambda E}^T$.

5. Transform back to X, i.e $\widehat{\boldsymbol{\mu}}_{RAWOGK} = \boldsymbol{D}\widehat{\boldsymbol{\mu}}(\boldsymbol{Y})$ and $\widehat{\boldsymbol{\Sigma}}_{RAWOGK} = \boldsymbol{D}\widehat{\boldsymbol{\Sigma}}(\boldsymbol{Y})\boldsymbol{D}^{\boldsymbol{T}}$.

Once these raw estimates are computed, they can be used to compute the robust Mahalanobis distances d_i = D($\boldsymbol{x}_i$, $\widehat{\boldsymbol{\mu}}_{RAWOGK}$, $\widehat{\boldsymbol{\Sigma}}_{RAWOGK}$) for all observations. If the Mahalanobis distance for any observation exceeds the critical value c = $\chi_p^2(0.9)$med (d_1, ... , d_n)/$\chi_p^2(0.5)$, it is declared an outlier. By using this cut off value and the robust Mahalanobis distance, a weight function can be defined and as in the FASTMCD algorithm the estimate is improved by a weighting step. The weighted estimates are denoted $\widehat{\boldsymbol{\mu}}_{OGK}$ and $\widehat{\boldsymbol{\Sigma}}_{OGK}$.

3.1.13. Comedian Approach

Sajesh and Srinivasan [63] proposed a method for the detection of outliers in multivariate data based on *comedian*, an alternative measure of dependence between two random variables introduced by Falk [64]. Let *X* and *Y* be two random variables then the *comedian* of *X* and *Y* is defined as

$$\mathrm{COM}(X,Y) = \mathrm{med}\,((X - \mathrm{med}(X))(Y - \mathrm{med}(Y))) \tag{3.10}$$

where med denotes median. It generalizes the Median Absolute Deviation (MAD) as it equals MAD^2 when $X = Y$ and also has the highest possible breakdown point [64]. Comedian parallels COV(*X*, *Y*), but COV(*X*, *Y*) requires the existence of the first two moments of *X* and *Y*, whereas COM(*X*, *Y*) always exists. The comedian is symmetric, location invariant and scale equivariant i.e., COM(*X*, *aY* + *b*) = *a* COM(*X*, *Y*) = *a* COM(*Y*, *X*). Hall and Welsh [65] discussed about the strong consistency and asymptotic normality of MAD. Falk [64] established similar results for *comedian*. In a similar way, a natural median based alternative to the coefficient of correlation is the *correlation median*

$$\delta(X,Y) = \delta = \frac{\mathrm{COM}(X,\ Y)}{\mathrm{MAD}(X)\mathrm{MAD}(Y)} \tag{3.11}$$

with δ Є [−1, 1] for bivariate data. Therefore *correlation median* of normal vectors (*X*, *Y*) as a measure of dependence between *X* and *Y* could very well be utilized [64].

Sajesh and Srinivasan [63] make use of the multivariate version of *comedian* estimate for the detection of outliers. Let **X** be an *n* × *p* data matrix with rows $\mathbf{x}_i^{\mathbf{T}}$ (*i* = 1, 2, …, *n*) and columns $\boldsymbol{X}_j$ (*j* = 1,2, …, *p*). Then the *comedian* matrix **COM**(**X**) is defined as

$$\mathbf{COM}(\mathbf{X}) = (\mathrm{COM}\,(\boldsymbol{X}i, \boldsymbol{X}j)), i, j = 1, 2, .., p. \tag{3.12}$$

Similarly, multivariate *correlation median* matrix **δ** is defined as,

$$\boldsymbol{\delta}(\mathbf{X}) = \boldsymbol{D}\mathbf{COM}(\mathbf{X})\boldsymbol{D}\mathrm{T} \tag{3.13}$$

where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $1/MAD(\mathbf{X}_i)$) (i = 1, …, p). Even though, these estimates have high breakdown values and some other properties, they have some drawbacks. In symmetry, med(X) = med(Y) = 0, med(X^2) = med (Y^2) = 1 which implies MAD(X) = MAD (Y) = 1, but $\Pr\{XY \leq 1\} = 0$ i.e., COM(X, Y) > 1. Thus *comedian* matrix **COM**(**X**), as a robust alternative to the covariance matrix, is in general not positive (semi-)definite [64]. The problem of non positive semi-definiteness of estimators frequently occurs in robust estimation of covariance matrix. Rousseeuw and Molenberghs [66] proposed several methods to deal with this problem. Maronna and Zamar [62] proposed a general method to obtain positive-definite and approximately affine equivariant robust scatter matrices. Sajesh and Srinivasan [63] adopted the following steps to overcome the non positive semi-definiteness of comedian matrix and to obtain robust estimates for location and scatter.

(i) Compute the eigen values λ_j and eigenvectors $\mathbf{e}_j$ of $\boldsymbol{\delta}(\mathbf{X})$ (j = 1, 2, …, p), and call **E** the matrix whose columns are the $\mathbf{e}_j$'s, so that $\boldsymbol{\delta}(\mathbf{X}) = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.

(ii) Let $\mathbf{Q} = \mathbf{D(X)}^{-1}\mathbf{E}$, where **D** is defined as above and $\mathbf{z}_i = \mathbf{Q}^{-1}\mathbf{x}_i$, i = 1, 2, …, n.

(iii) The resulting robust estimates for location ($\boldsymbol{m}(\mathbf{X})$) and scatter ($\boldsymbol{S}(\mathbf{X})$) is then defined as

$$\boldsymbol{S}(\mathbf{X}) = \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}\mathbf{T} \text{ and } \boldsymbol{m}(\mathbf{X}) = \mathbf{Q}\boldsymbol{l} \tag{3.14}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}(\mathrm{MAD}(\boldsymbol{Z}_1)^2, \ldots, \mathrm{MAD}(\boldsymbol{Z}_p)^2)$ and $\boldsymbol{l} = (\mathrm{med}(\boldsymbol{Z}_1), \ldots, \mathrm{med}(\boldsymbol{Z}_1))^{\boldsymbol{T}}$.

The estimates can be improved through an iterative process, by replacing $\boldsymbol{\delta}$ with $\boldsymbol{S}$ and repeat the steps (i), (ii) and (iii). The primary interest is the detection of outliers, by using a robust Mahalanobis distance defined as,

$$RD(\boldsymbol{xi}, \boldsymbol{m}) = rdi = (\boldsymbol{xi} - \boldsymbol{m})T\,\boldsymbol{S} - \mathbf{1}\,(\boldsymbol{xi} - \boldsymbol{m}), i = 1,2, \ldots, n, \tag{3.15}$$

where $\boldsymbol{S}$ and $\boldsymbol{m}$ are defined in (3.14).

The efficacy of the detection of outliers could be considered by a suitable cut off value *cv* defined as

$$cv = 1.4826\,\frac{\chi^2_{\mathrm{p}}(0.95)\mathrm{median}\,(\mathrm{rd}_1, \ldots, rd_n)}{\chi^2_{\mathrm{p}}(0.5)} \tag{3.16}$$

Accordingly, if any $RD(\boldsymbol{x_i}, \boldsymbol{m}) > cv$, the corresponding observation $\boldsymbol{x}_i$ can be considered as outlier. By using this cut off value and the robust Mahalanobis distance, a weight function can be defined and robust estimates for location and

scatter can be obtained. These estimates are positive definite and approximately affine equivariant. In addition, the estimates obtained by *comedian* method would have high-breakdown value and helps detection of large cluster of outliers. The efficiency of the method increases with the increase in the dimension of datasets as examined through various numerical studies.

3.1.14. Other Distance-based Methods

Oyeyemi and Ipinyomi [67], proposed a robust method of estimating a covariance matrix in a multivariate data set. The proposed robust method performs favourably well in the detection of single or fewer outliers especially for small sample size and when the magnitude of outliers is relatively small. Outlier detection on time series data plays an important role in life. Ren *et al* [68] proposed a method of outlier detection on time series data mainly aiming at the multivariate type. The improved ant colony algorithm is used for data clustering in classification of time series data. Both the distance of inner-clusters and inter-clusters are considered to ensure the accuracy of the clustering. The objects which have significant changes from the neighbors are identified as outliers. The presence of missing values is more a rule than an exception in business surveys and poses additional severe challenges to the outlier detection. Todorov *et al* [69] compared some multivariate outlier detection methods which can cope with incomplete data through a simulation study and identified methods finding the outliers with low false discovery rate. Rousseeuw and Hubert [70] presented an overview of several robust methods and outlier detection tools suitable for univariate, low-dimensional, and high-dimensional data such as estimation of location and scatter, linear regression, principal component analysis, and classification.

3.2. Non-Traditional Methods

A common limitation with all robust distance-based outlier detection methods is the requirement to find a subset of outlier-free data from which robust estimates of the mean vector and covariance matrix can be obtained. Unfortunately, there is no existing method that can find an outlier-free subset with 100% certainty. In other words, there is always a chance that the “outlier-free” sample contains some outliers. Researchers have proposed alternative non-traditional outlier detection methods that attempt to avoid robust Mahalanobis distances altogether. In the following paragraphs, the significant non-traditional outlier detection methods found in the technical literature are outlined. As in the previous section, these methods are discussed in chronological order to illustrate how these methods have evolved over time.

3.2.1. Principal Component Methods

One of the earliest distance-free methods for detecting multiple outliers in multivariate data is described by Gnanadesikan and Kettenring [25] and is originally attributed to Rao [71]. This method makes the assumption that the dataset falls in the linear subspace defined by the first p - q principal components of the sample covariance matrix. Under this assumption, it is argued that outliers will have a large deviation from this sub-space as measured by the sum of the magnitudes of their projections onto the last q eigenvectors. More specifically, outliers in a $n \times p$ dataset, $\mathbf{Y}$, are observations, $\mathbf{y}_j$, with large values of:

$$d_j^2 = \sum_{i=p-q+1}^{p} \left[\boldsymbol{I}_i^T (\boldsymbol{y}_j - \bar{\boldsymbol{y}}) \right] \tag{3.17}$$

where $\boldsymbol{I}_i$ = the eigenvector corresponding to the i^{th} smallest eigen value of the covariance matrix and $\bar{\boldsymbol{y}}$ = mean vector of $\mathbf{Y}$.

Gnanadesikan and Kettenring [25] suggest analyzing the d_j^2 values through the use of a gamma probability plot where the shape parameter is estimated using a method proposed by Wilk and Gnanadesikan [72]. In addition to Rao's method, Gnanadesikan and Kettenring suggest other informal uses of the principal component scores for detecting outliers. Unfortunately, a limitation of these methods is they are devoid of any formal tests of significance, relying upon the analyst to subjectively determine how an outlier should manifest itself.

3.2.2. Mahalanobis Distance Decomposition Method

As an alternative to computing robust Mahalanobis distances to detect outliers, Kim [73], derives two decompositions of the Mahalanobis distance and uses scatter plots of the component terms to uncover outlying observations. Thus, rather than using the Mahalanobis distances themselves to find outliers, Kim suggests analyzing the constituent parts of the Mahalanobis distances for an observation to determine how the distance was achieved. Kim provides no guidance on the distribution of the components of the Mahalanobis distance, thus requiring subjective analysis of the suggested scatter plots to identify outliers.

3.2.3. Projection Pursuit Detection

In order to avoid the masking and swamping effects associated with the classical Mahalanobis distance detector as well as the computational complexities of robust distance detection methods, Pan *et al*. [74], proposed a method that uses univariate projections of the original data and univariate outlier detection to identify

multivariate outliers. The method begins by projecting the original data onto a vector located on the *p*-dimensional unit hypersphere. Based on tests with relatively small datasets, Pan *et al*. demonstrated that their method is effective at detecting outliers while achieving relatively low false alarm rates. No evidence is provided to suggest this method is scalable to larger problems. In fact, using this method for high dimensional datasets can be problematic since the number of projection vectors generated to achieve uniform coverage of the *p*-dimensional unit hypersphere grows non-linearly with *p*. Further discussion of this problem is provided by Fang and Wang [80][75].

3.2.4. Juan-Prieto Method

Empirical tests conducted by Juan and Prieto[76], indicate that the robust distance methods of Rocke and Woodruff [50], Hawkins[49], Rousseeuw and Van Driessen [43], and Maronna and Yohai [77], have difficulty in detecting clusters of concentrated outliers, particularly when the clusters are relatively close to the good data. To overcome this perceived weakness of robust distance methods, Juan and Prieto suggested a distance free method based on angles. Specifically, the authors state that the projections of observations on the *p*-dimensional unit hypersphere are uniformly distributed when the observations have an ellipsoidal distribution, as shown by Eaton [78]. Based on this characteristic, Juan and Prieto claimed that the angles between the projected observation vectors and an arbitrary reference direction, $\mathbf{u}_0$, have a Beta distribution. The form of the Beta distribution is provided by the authors.

To detect outliers, the original observations are projected onto the unit hypersphere. A reference direction, $\mathbf{u}_0$, is then selected using a method suggested by Juan and Prieto [76]. The angles between the projected observations and $\mathbf{u}_0$ are then computed. The authors then suggest using a *QQ* - plot of the angles to determine if they follow the beta distribution. Alternatively, the distributional fit of the angles can be assessed by analyzing the spacings between the ordered values of $F(\mathbf{w}_i)$, the theoretical distribution function of the angles evaluated at each angle, $\mathbf{w}_i$. If the angles actually follow the prescribed distribution, the spacing should be uniformly distributed. To test this hypothesis, Juan and Prieto suggested using the distribution of the largest spacing in a uniform sample introduced by David [79]. From this distribution function, a critical value for the largest spacing can be computed and all the largest spacing tested for significance. If the test fails, any corresponding observations preceding the largest spacing are considered outliers. To detect multiple outlier clusters, this entire process is repeated until the spacing indicates uniformity of the angles.

3.2.5. Chiang-Pell-Seasholtz PCA Method

Where Gnanadesikan and Kettenring [25], proposed somewhat informal methods for using PCA to find multivariate outliers, Chiang *et al*. [80], presented a PCA method that included significance tests for outliers. The method began by performing a PCA on the original data to arrive at the $p \times a$ matrix, $\mathbf{P}$, containing the eigenvectors corresponding to the a largest eigen values. In addition to testing if an observation is an outlier using the components for the a largest eigen values, Chiang *et al*. also suggest testing the observation using the $p - a$ components for the remaining eigen values. To perform this test, the authors recommend using the Q-statistic of Jackson and Mudholkar [81]. The threshold value for the Q-statistic is provided by Chiang *et al*. If the Q-statistics for an observation exceed their respective critical value, the observation is labeled an outlier and removed from the data set. Once all observations are tested, the entire process is repeated using only the non-outlying observations. The algorithm terminates when no additional observations are labeled outliers between iterations, or when the total number of outliers detected reaches $n/2$.

3.2.6. Max-Eigen Difference (MED) Method

Adding to the arsenal of principal component-based outlier detection methods, Gao *et al.* [82], proposed the Max-Eigen Difference (MED) method. The method proceeds by computing the eigen values and eigenvectors of the sample covariance matrix of the entire dataset. For each observation, $\boldsymbol{x}_i$, the eigen values and eigenvectors are then computed for the covariance matrix obtained when $\boldsymbol{x}_i$ is removed from the dataset.

Gao *et al*. [82], demonstrated that large MED values indicate outlier observations. Specifically, the decomposition illustrates that an observation with a large MED may indicate: *i*) the observation has a first principal component score that is much larger than the other observations; *ii*) the observation may have relatively large scores on the other component axes; and *iii*) the observation is not close to the centroid of the data. An observation with large MED may possess any combination of these characteristics. Based on the properties of the MED, Gao *et al*. [82], recommended detecting outliers by plotting the MED values against the observation indices. Any observations that appear to have a large MED relative to the other observations are labeled as outliers. This labeling is a subjective decision made by the analyst.

3.2.7. Other Non-Traditional Methods

Singh *et al*. [83], have proposed an unsupervised clustering scheme for isolating atypical behaviors, a parameter less outlier detection method based on wavelets and a new feature for characterizing intrusions based on the repetition of an intrusion attempt from one system to another. Al-Zoubi [84], discussed a method based on clustering approaches for outlier detection using PAM clustering algorithm. Ueda [85], presented a simple and efficient method to detect multiple outliers using a modification of the Akaike's Information Criterion. It is well known that if a multivariate outlier has one or more missing component values, then *multiple imputation* (MI) methods tend to impute non-extreme values and make the outlier become less extreme and less likely to be detected. Dang and Serfling [86] proposed nonparametric depth-based multivariate outlier identifiers for such type of data. Two criteria, an 'outlier recovery probability' and a 'relative accuracy measure', are developed, based on depth functions. Yan [87] proposed a novel method integrating self-organizing map (SOM) with adaptive non-linear map (ANLM) to facilitate visualizing and detecting outliers in high dimensional complex data.

It is interesting to note that volume of data collected is getting exponentially increased by the day with the availability of powerful storage devices. It is equally important to analyze the data and draw meaningful inference because of its impact on various applications, like the collection and analysis of genomic data is important in the study of human diseases and drug discovery. In this process, detection of multiple outliers in multidimensional data is crucial for its influence in the inferential process and its interpretation in wide variety of applications.

4. Comparative Study

Since the comparison of all the above mentioned methods is tedious, four recent and important methods namely, FAST-MCD, Kurtosis, OGK and Comedian are selected. Performance of these methods is evaluated through simulation using the parameters of *Success Rate* (SR) and *False Detection Rate* (FDR). While success rate measures the detection of true outliers, FDR appraise the false detection of normal observations as outliers. In other words, success rate is a measure of masking as it reveals the number of true outliers which are not detected and false detection rate is a measure of swamping as it shows the number of inliers detected as outliers.

For a given contamination level α, a set of $100(1-\alpha)$ observations from an $N(\boldsymbol{0}, \boldsymbol{I})$ distribution with dimension p has been generated and 100α additional observations are added from a $N(\xi\boldsymbol{u}, \lambda\boldsymbol{I})$ distribution, where $\boldsymbol{u}$ denotes the vector $(1, 1, \ldots, 1)^T$. This experiment has been conducted for different values of the sample space dimension p(p = 5, 10, 20) and the contamination level α (α = 0.1, 0.2, 0.3).

To check the efficiency of detecting very small deviations the experiment has been conducted for small values of ξ (ξ = 5, 10) and λ (λ = 0.01, 0.25, 1). For each set of values, 100 samples have been generated and parameters are estimated. To attain 50% of breakdown value for FAST-MCD method the subset size $h = [(n+p+1)/2]$ is used. The scale estimate Q_n proposed by Croux and Rousseeuw[88] is used as initial scale estimates for OGK. MATLAB codes have been used for Kurtosis, FAST-MCD and Comedian methods and OGK is available in R package.

Table 3.1 presents the success rates of Comedian, Kurtosis, FAST-MCD and OGK methods for each set of the parameter values. For ξ = 5, except in two situations (for p = 5, α = 0.3, λ = 0.01 and p = 10, α = 0.3, λ = 0.01) Comedian method attains 95% or more success rate. Sample-space dimension p has significant influence on outlier detection methods. It is interesting to note that, for each value of α and λ success rate of the Comedian increases with increasing dimension. Although, the success rate of OGK also increases with dimension, the rate of increase is low compared to that of Comedian. Regarding the success rate, Kurtosis seems not much influenced by the sample-space dimension. But, this is not the case of FAST-MCD. Success rate of FAST-MCD decreases with increasing dimension. For example, for α = 0.3, λ = 0.25 and ξ = 5 the success rate of FAST-MCD is decreases from 60% for p = 5 to 41% for p = 10 and 0% for p = 20.

Regarding the value of λ, Comedian performs consistently well for all values except for two cases where the method scored less than 95% success rate. It is important to note that, for λ = 1 the success rate of Kurtosis method decreases with increasing dimension. For λ = 1, the success rate of Kurtosis method is 97% for p = 5, 75% for p = 10 and 49% for p = 20. Peña and Prieto [61] also shows similar results and states that, this case tends to be one of the most difficult ones for the kurtosis algorithm because the objective function is nearly constant for all directions, and for finite samples, it tends to present many local minimizers, particularly along directions that are nearly orthogonal to the outliers. It is just opposite in the case of FAST-MCD method that, for ξ = 5, only cases where it attains more than 95% success rate is when λ = 1. Its behavior is worse for both the remaining values of λ. Unlike FAST-MCD and Kurtosis, OGK performs almost steadily for all values of λ.

Table 3.1: Success rates of outlier detection methods

			$\xi = 5$				$\xi = 10$			
p	***λ***	***α***	***Comedian***	***Kurtosis***	***FAST-MCD***	***OGK***	***Comedian***	***Kurtosis***	***FAST-MCD***	***OGK***
5	0.01	0.1	100	100	100	100	100	100	100	100
		0.2	100	99	39	100	100	100	100	100
		0.3	70	99	0	34	100	100	0	100
	0.25	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	100	100	100	100	100	100
		0.3	95	98	60	81	100	100	100	100
	1	0.1	100	100	100	100	100	46	100	100
		0.2	100	98	100	100	100	1	100	100
		0.3	99	97	100	83	100	0	2	100
10	0.01	0.1	100	100	100	100	100	100	100	100
		0.2	100	99	0	100	100	100	0	100
		0.3	83	91	0	38	100	92	0	100
	0.25	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	41	100	100	100	100	100
		0.3	99	79	0	99	100	90	0	100
	1	0.1	100	100	100	100	100	100	100	100
		0.2	100	75	100	100	100	100	100	100
		0.3	100	21	99	100	100	38	100	100
20	0.01	0.1	100	100	0	100	100	100	86	100
		0.2	100	85	0	100	100	94	0	100
		0.3	99	0	0	52	100	3	0	100
	0.25	0.1	100	100	95	100	100	100	100	100
		0.2	100	90	0	100	100	92	0	100
		0.3	100	3	0	100	100	2	0	100
	1	0.1	100	49	100	100	100	46	100	100
		0.2	100	1	100	100	100	1	100	100
		0.3	100	0	2	100	100	0	2	100

Table 3.2: False Detection Rates of outlier detection methods

			$\xi = 5$				$\xi = 10$			
λ	p	α	*Comedian*	*Kurtosis*	*FAST-MCD*	*OGK*	*Comedian*	*Kurtosis*	*FAST-MCD*	*OGK*
0.01	5	0.1	4	7	17	15	3	9	18	11
		0.2	5	7	41	10	6	7	12	12
		0.3	4	5	51	10	4	5	47	7
	10	0.1	4	6	22	16	2	7	23	19
		0.2	6	42	44	13	5	5	42	11
		0.3	3	45	54	7	5	45	56	9
	20	0.1	1	10	39	18	1	8	38	16
		0.2	3	40	47	12	3	40	47	14
		0.3	3	40	63	12	3	40	61	7
0.25	5	0.1	3	5	18	16	5	7	13	14
		0.2	2	5	11	11	3	6	12	10
		0.3	2	5	32	7	2	5	8	6
	10	0.1	2	5	24	20	2	8	21	17
		0.2	2	7	36	10	2	7	15	13
		0.3	2	31	40	9	4	24	37	7
	20	0.1	1	9	38	16	1	8	28	19
		0.2	2	13	39	10	1	14	39	12
		0.3	1	39	40	7	1	40	41	9
1	5	0.1	3	6	14	15	3	6	19	14
		0.2	2	6	9	13	2	6	10	11
		0.3	2	6	7	7	2	6	7	8
	10	0.1	2	9	23	16	2	6	23	18
		0.2	2	6	15	13	3	6	18	10
		0.3	2	6	13	7	1	7	9	7
	20	0.1	2	8	28	16	2	9	28	14
		0.2	1	5	18	12	1	5	19	11
		0.3	1	4	27	8	1	4	30	7

Amount of contamination α is an important parameter to be analyzed because most of the outlier detection methods fall short when there is large amount of contamination. Comedian achieved 95% or more success rates for all values of α, except in two cases. Regarding the amount of contamination, Kurtosis method seems to perform better for $p = 5$, while its behavior is worse for large values of sample-space dimension. In two cases of $p = 5$; $\lambda = 1$and $p = 10$; $\lambda = 1$ alone FAST-

MCD detects 30% of contamination and failed to do so in the rest of cases. The worst case is for $\xi = 5$, $\lambda = 0.01$ and $\alpha = 0.3$ and for all values of p FAST-MCD has 0% success rate. Even though OGK performs better than FAST-MCD and Kurtosis, OGK fails to achieve at least 95% of success rate for any of the cases with $\xi = 5$, $\lambda = 0.01$ and $\alpha = 0.3$. For $\xi = 10$, Comedian and OGK attain optimal success rates in all cases, while the success rates of FAST-MCD and Kurtosis follow the same pattern of $\xi = 5$.

False detection rate is also an important property to be examined for comparison of various methods. Table 3.2 gives the false detection rates of Comedian, Kurtosis, FAST-MCD and OGK methods from 100 independent samples. Based on the success rates, Comedian and OGK methods provide almost equal results. But a comparison based on FDR will support the supremacy of Comedian among other similar methods. It is clear that in every situation Comedian possesses much lower false detection rates than other methods and also it reduces with increasing dimension. Also, the maximum false detection rate is 6 for Comedian, 45 for Kurtosis, 63 for FAST-MCD and 20 for OGK, based on all possible situations.

5. Conclusion

Successful identification of outliers has a very close connection with robust estimation. Classical estimators such as the mean and covariance matrix are not suitable for data containing outliers and can cause the statistical analysis to produce results exactly opposite to the correct conclusions. Thus, robust statistical techniques that can be computed in a reasonable time should be used if outliers are thought to be present [93]. Within the field of statistics, there are two broad approaches to outlier identification. Distance-based methods, such as MCD, *Comedian* and BACON, are based on obtaining robust estimates of the mean and covariance matrix so that a robust Mahalanobis distance can be computed for each point. Promising avenues for future research include finding robust covariance estimates that can be quickly computed, while still maintaining robustness against outliers in a variety of configurations. Non-traditional methods aim to find the best projections that reveal the outliers in a highly visible placement. Such approaches can find outliers in a wide variety of configurations since the original placement of outliers is transformed to more informative projections. However, such methods tend to be very computationally intensive and are not currently suitable for large datasets. It remains to be seen whether computationally efficient methods of projection pursuit can be found to enable this strategy to be used in data-mining and similar applications.

Performances of four recent methods are evaluated through simulation using the parameters of Success Rate and False Detection Rate. The simulation study has

explored and examined almost all possible situations by varying parameters. Results show that, when compared to other methods, the comedian method is able to detect all the outliers efficiently based on success rate and false detection rate. Also the efficiency of comedian method increases with the increase in the dimension of data.

Acknowledgement

The authors would like to thank Naval Research Board, DRDO, New Delhi, India, for providing the grants in carrying out the above research work

References

1. Barnett, V. and Lewis, T. Outliers in Statistical Data. John Wiley and Sons, Chichester, England, 1994. DOI:10.1016/0169-2070(95)00625-7
2. Grubbs, F.E. Procedures for Detecting Outlying Observations in Samples. Technometrics, 11: 1-21, 1969. DOI: 10.1080/00401706.1969.10490657
3. Hawkins, D.M. Identification of Outliers. Chapman and Hall, London, 1980. DOI: 10.1002/bimj.4710290215
4. Kafadar, K. and Morris, M. Data-based Detection of Potential Terrorist Attacks. White Paper, 2002.
5. Banks, D. Statistics for Homeland Defense. Chance, 15: 8-10, 2002.
6. Huber, P.J. Robust Statistics. Wiley, New York, 1981.
7. Tukey, J.W. Exploratory Data Analysis. Addison-Wesley, New York, USA, 1977.
8. Hoaglin, D.C., Iglewicz, B. and Tukey, J.W. Performance of Some Resistant Rules for Outlier Labeling. Journal of the American Statistical Association, 81: 991- 999, 1986. DOI:10.2307/2289073
9. Hoaglin, D.C. and Iglewicz, B. Fine Tuning Some Resistant Rules for Outlier Labeling. Journal of the American Statistical Association, 82: 1147-1149, 1987. DOI: 10.1080/01621459.1987.10478551
10. Banerjee, S. and Iglewicz, B. A Simple Univariate Outlier Identification Procedure Designed for Large Samples. Communications in Statistics - Simulation and Computation, 36: 249-263, 2007. DOI: 10.1080/03610910601161264
11. Kimber, A.C. Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions. Applied Statistics, 39: 21–30, 1990.

12. Van der Loo, M.P.J. Distribution Based Outlier Detection in Univariate Data, Statistics Netherlands, 2010.

13. Gentleman, J.F. and Wilk, M.B. Detecting outliers in a two-way table: I. Statistical behaviour of residuals. Technometrics, 17: 1–14, 1975. DOI: 10.1080/00401706.1975.1048926

14. Gentleman, J.F. and Wilk, M.B. Detecting outliers: II. Supplementing the direct analysis of residuals. Biometrics, 31: 387–410, 1975.

15. John, J.A. and Draper, N.R. On testing for two or one outliers in two-way tables. Technometrics, 20: 69–78, 1978. DOI: 10.1080/00401706.1978.10489618

16. Prescott, P. An approximate test for outliers in linear models. Technometrics, 17: 129–132, 1975. DOI: 10.1080/00401706.1975.10489282

17. John, J.A. Outliers in factorial experiments. Applied Statistics, 27: 111–119, 1978.

18. Balasooriya, U., Tse, Y.K. and Liew, Y.S. An empirical comparison of some statistics for identifying outliers and influential observations in linear regression models. Journal of Applied Statistics, 14: 177–184, 1987. DOI: 10.1080/02664768700000022

19. Tietjen, G.L., Moore, R.H. and Beckman, R.J. Testing for a single outlier in simple linear regression. Technometrics, 15: 717–721, 1973. DOI: 10.1080/00401706.1973.10489106

20. Andrews, D.F. and Pregibon, D. Finding the outliers that matter. Journal of the Royal Statistical Society, Series B, 40: 85–93, 1978.

21. Cook, R.D. and Weisberg, S. Residuals and Influence in Regression. Chapman and Hall, NewYork, USA, 1982.

22. Cook, R.D. Detection of influential observations in linear regression. Technometrics, 19: 15–18, 1977. DOI:10.2307/1268249

23. Draper, N.R. and John, J.A. Influential observations and outliers in regression. Technometrics, 23: 21–26, 1981. DOI: 10.1080/00401706.1981.10486232

24. Balasooriya, U. and Tse, Y.K. Outlier detection in linear models: a comparative study in simple linear regression. Communications in Statistics – Theory and Methods, 15: 3589–3597, 1986. DOI: 10.1080/03610928608829332

25. Gnanadesikan, R. and Kettenring, J.R. Robust Estimates, Residuals, and Outlier Detection with Multi-response Data. Biometrics, 28: 81-124, 1972.

26. Beckman, R.J. and Cook, R.D. Outlier...s. Technometrics, 25: 119-163, 1983.

27. Hodges, J.L. Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

28. Hampel, F.R. Contributions to the Theory of Robust Estimation. Ph.D Thesis, University of California, Berkeley, CA, 1968.

29. Hampel, F.R. A Generalized Qualitative Definition of Robustness. Annals of Mathematical Statistics, 42: 1887-1896, 1971. DOI:10.1214/aoms/1177693054

30. Lopuhaä, H.P. and Rousseeuw, P.J. Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. The Annals of Statistics, 19: 229-248, 1991. DOI:10.1214/aos/1176347978

31. Rousseeuw, P.J. and Leroy, A.M. Robust Regression and Outlier Detection. John Wiley and Sons, New York, USA, 1987. DOI: 10.1002/cem.1180020410

32. Donoho, D.L. and Huber, P.J. 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, 1983.

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

34. Hubert, M., Rousseeuw, P.J. and Van Aelst, S. High-Breakdown Robust Multivariate Methods. Statistical Science, 23: 92-119, 2008. DOI:10.1214/088342307000000087

35. Becker, C. and Gather, U. The Masking Breakdown Point of Multivariate Outlier Identification Rules. Journal of the American Statistical Association, 94: 947-955, 1999. DOI:10.2307/2670009

36. Hadi, A.S. Identifying Multiple Outliers in Multivariate Data. Journal of the Royal Statistical Society. Series B (Methodological), 54: 761-771, 1992.

37. Campbell, N.A. Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation. Applied Statistics, 29: 231-237, 1980.

38. Maronna, R.A. Robust M-Estimators of Multivariate Location and Scatter. The Annals of Statistics, 4: 51-67, 1976. DOI:10.1214/aos/1176343347

39. Rousseeuw, P.J. Multivariate Estimation with High Breakdown Point. Fourth Pannonian Symposium on Mathematical Statistics and Probability, 1983.

40. Rousseeuw, P.J. and van Zomeren, B.C. Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association, 85: 633-639, 1990. DOI: 10.1080/01621459.1990.10474920

41. Butler, R.W., Davies, P.L. and Jhun, M. Asymptotics for the Minimum Covariance Determinant Estimator. The Annals of Statistics, 21: 1385-1400, 1993. DOI:10.1214/aos/1176349264

42. Davies, L. The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator. The Annals of Statistics, 20: 1828-1843, 1992. DOI:10.1214/aos/1176348891

43. Rousseeuw, P.J. and Van Driessen, K. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics, 41: 212-223, 1999.

44. Stahel, W.A. Robuste Schatzungen: Infinitesimale Optimalitat und Schatzungen von Kovarianzmatrizen, Ph.D Thesis. ETH Zurich, Zurich, Switzerland, 1981.

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

46. Gasko, M. and Donoho, D.L. Influential Observation in Data Analysis. American Statistical Association Proceedings of the Business and Economic Statistics Section, 1: 104-109, 1982.

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

48. Atkinson, A.C. Fast Very Robust Methods for the Detection of Multiple Outliers. Journal of the American Statistical Association, 89: 1329-1339, 1994. DOI: 10.1080/01621459.1994.10476872

49. Hawkins, D.M. The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data. Computational Statistics and Data Analysis, 17: 197-210, 1994. DOI: 10.1016/0167-9473(92)00071-X

50. Rocke, D.M. and Woodruff, D.L. Identification of Outliers in Multivariate Data. Journal of the American Statistical Association, 91: 1047-1061, 1996. DOI:10.2307/2291724

51. Rocke, D.M. and Woodruff, D.L. Computation of Robust Estimates of Multivariate Location and Shape. Statistica Neerlandica, 47: 27-42, 1993. DOI: 10.1111/j.1467-9574.1993.tb01404.x

52. Woodruff, D.L. and Rocke, D.M. Heuristic Search Algorithms for the Minimum Volume Ellipsoid. Journal of Computational and Graphical Statistics, 2: 69-95, 1993. DOI:10.1080/10618600.1993.10474600

53. Woodruff, D.L. and Rocke, D.M. Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators. Journal of the American Statistical Association, 89: 888-896, 1994. DOI:10.2307/2290913

54. Rocke, D.M. Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension. The Annals of Statistics, 24: 1327-1345, 1996. DOI:10.1214/aos/1032526972

55. Egan, W.J. and Morgan, S.L. Outlier Detection in Multivariate Analytical Chemical Data. Analytical Chemistry, 70: 2372-2379, 1998.

56. Zani, S., Riani, M. and Corbellini, A. Robust Bivariate Boxplots and Multiple Outlier Detection. Computational Statistics and Data Analysis, 28: 257-270, 1998.

57. Bebbington, A.C. A Method of Bivariate Trimming for Robust Estimation of the Correlation Coefficient. Applied Statistics, 27: 221-226, 1978.

58. Ammeraal, L. Programming Principles in Computer Graphics. Wiley, New York, USA, 1992.

59. Hadi, A.S. A Modification of a Method for the Detection of Outliers in Multivariate Samples. Journal of the Royal Statistical Society, Series B, 56: 393-396, 1994.

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

61. Peña, D. and Prieto, F.J. Multivariate Outlier Detection and Robust Covariance Matrix Estimation. Technometrics, 43: 286 – 300, 2001.

62. Maronna, R.A. and Zamar, R.H. Robust Estimates of Location and Dispersion for High-Dimensional Datasets. Technometrics, 44: 307-317, 2002. DOI:10.1198/004017002188618509

63. Sajesh, T.A. and Srinivasan, M.R. Outlier Detection for High Dimensional Data using Comedian Approach. Journal of Statistical Computation and Simulation, 82(5): 745-757, 2012. DOI:10.1080/00949655.2011.552504

64. Falk, M. On MAD and Comedian. Annals of the Institute of Statistical Mathematics, 49: 615-644, 1997.

65. Hall, P. and Welsh, A.H. Limit Theorems for the Median Deviation. Annals of Institute of Statistical Mathematics, 37: 27-36, 1985. DOI:10.1007/BF02481078

66. Rousseeuw, P.J. and Molenberghs, G. Transformation of Non-Positive Semi-Definite Correlation Matrices. Communications in Statistics, Part A-Theory and Methods. 22: 965-984, 1993.

67. DOI:10.1080/03610928308831068

68. Oyeyemi, G.M. and Ipinyomi, R.A. A Robust Method of Estimating Covariance Matrix in Multivariate Data Analysis. African Journal of Mathematics and Computer Science, 3(1): 1-18, 2010.

69. Ren, J., Li, H., Hu, C. and He, H. ODMC: Outlier Detection on Multivariate Time Series Data based on Clustering. Journal of Convergence Information Technology, 6(2): 70-77, 2011.

70. Todorov, V., Templ, M. and Filzmoser, P. Detection of multivariate outliers in business survey data with incomplete information. Advances in Data Analysis and Classification, 5: 37–56, 2011. DOI: 10.1007/s11634-010-0075-2

71. Rousseeuw, P.J. and Hubert, M. Robust statistics for outlier detection. WIREs Data Mining and Knowledge Discovery, 1: 73–79, 2011. DOI: 10.1002/widm.2

72. Rao, C.R. The Use and Interpretation of Principle Component Analysis in Applied Research. Sankhya, 26: 329-358, 1964.

73. Wilk, M.B. and Gnanadesikan, R. Graphical Methods for Internal Comparisons in Multi-response Experiments. Annals of Mathematical Statistics, 35: 613-631, 1964.

74. Kim, M.G. Multivariate Outliers and Decompositions of Mahalanobis Distance. Communications in Statistics-Theory and Methods, 29: 1511-1526, 2000. DOI: 10.1080/03610920008832559

75. Pan, J.X., Fung, W.K. and Fang, K.T. Multiple Outlier Detection in Multivariate Data Using Projection Pursuit Techniques. Journal of Statistical Planning and Inference, 83: 153-167, 2000.

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

77. Juan, J. and Prieto, F.J. Using Angles to Identify Concentrated Multivariate Outliers. Journal of the American Statistical Association, 43: 311-322, 2001.

78. Maronna, R.A. and Yohai, V.J. The Behavior of the Stahel-Donoho Robust Multivariate Estimator. Journal of the American Statistical Association, 90: 330-341, 1995. DOI:10.2307/2291158

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

80. David, H.A. Order Statistics. Wiley, New York, 1981.

81. Chiang, L.H., Pell, R.J. and Seasholtz, M.B. Exploring Process Data with the Use of Robust Outlier Detection Algorithms. Journal of Process Control, 13: 437- 449, 2003. DOI: 10.1016/S0959-1524(02)00068-9

82. Jackson, J.E. and Mudholkar, G.S. Control Procedures for Residuals Associated with Principal Component Analysis. Technometrics, 21: 341-349, 1979. DOI: 10.1080/00401706.1979.10489779

83. Gao, S., Li, G. and Wang, D. A New Approach for Detecting Multivariate Outliers. Communications in Statistics-Theory and Methods, 34: 1857-1865, 2005. DOI: 10.1081/STA-200066315

84. Singh, G., Massenglia, F., Fiot, C., Marascu, A. and Poncelet, P. Data Mining for Intrusion Detection: from Outliers to True Intrusions. The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD–09), 891-898, 2009. DOI: 10.1007/978-3-642-01307-2_93

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

86. Ueda, T. A Simple Method for the Detection of Outliers. Electronic Journal of Applied Statistical Analysis, 1: 67-76, 2009. DOI: 10.1285/i20705948v2n1p67

87. Dang, X. and Serfling, R. A numerical study of multiple imputation methods using nonparametric multivariate outlier identifiers and depth-based performance criteria with clinical laboratory data. Journal of Statistical Computation and Simulation, 81(5): 547–560, 2011. DOI: 10.1080/00949650903437842

88. Yan, X. Multivariate outlier detection based on self-organizing map and adaptive nonlinear map and its application. Chemometrics and Intelligent Laboratory Systems, 107: 251–257, 2011. DOI:10.1016/j.chemolab.2011.04.007

89. Croux, C. and Rousseeuw, P. J. Time-efficient algorithms for two highly robust estimators of scale. Computational Statistics, 2: 411-428, 1992.