A STUDY ON ROBUST MULTIVARIATE TECHNIQUES

Submitted to

University of Calicut

for the award of the degree of

DOCTOR OF PHILOSOPHY

IN STATISTICS

Under the faculty of Science

by

SAJANA, O.K.

Under the guidance of

Dr SAJESH T.A

POSTGRADUATE AND RESEARCH

DEPARTMENT OF STATISTICS

ST THOMAS' COLLEGE (AUTONOMOUS)

THRISSUR, KERALA-680001

INDIA

MARCH-2020

DEPARTMENT OF STATISTICS
ST. THOMAS' COLLEGE (Autonomous)
THRISSUR, KERALA - 680001, INDIA.
www.stthomas.ac.in/stat.htm

Dr. Sajesh T.A, M.Phil, Ph.D.
Assistant Professor of Statistics
Ph: +919747936823
E mail: sajesh.t.abraham@gmail.com

23-03-2020

CERTIFICATE

This is to certify that the thesis entitled "**A STUDY ON ROBUST MULTIVARIATE TECHNIQUES**", is an authentic record of research work carried out by **Ms. Sajana O.K** under the supervision in fulfilment of the requirement for the degree of Doctor of Philosophy, in Statistics of University of Calicut. The results embodied in this thesis have not been included in any other thesis submitted previously for the award of any degree or diploma of any other university or institution. Also certified that the contents of the thesis have been checked using anti-plagiarism data base and no unacceptable similarity was found through the software check.

Dr. SAJESH T.A

(Research Guide)

DECLARATION

I here by declare that the thesis entitled **"A STUDY ON ROBUST MULTIVARIATE TECHNIQUES"**, submitted to the University of Calicut, for the award of the degree of **Doctor of Philosophy in Statistics**, under the faculty of Science is an bonafide work done by me under the supervision of **Dr Sajesh T.A**, Assistant Professor, Department of Statistics, St Thomas' College(Autonomous), Thrissur, Kerala.

I also declare that this thesis contains no material which has been accepted for the award of any other degree or diploma of any University or institution and to the best of my knowledge and belief, it contains no material previously published by any other person, except where due reference made in the text of the thesis.

SAJANA O.K

AKNOWLEDGEMENTS

First and foremost, I thank God, Almighty for the boundless blessings, which led me to fulfill my ambition.

At this moment of accomplishment I would like to express thanks to my research supervisor and guide Dr Sajesh T.A, who accepted me as his Ph.D student and offered me his mentorship. This work would not have been possible without his guidance, motivation and encouragement.

Besides my supervisor, I would like to thank Dr Chacko V.M, Head, Department of Statistics, Dr. Rani Sebastian, Dr. Jeena Joseph and Dr. Nicy Sebastian for their insightful comments and constant advice.

Let me thank my family members, Amma, Achan, Brothers and Sister-in-laws for motivating me from the registration of my PhD to this date of submission in all dimensions, especially to my elder sister-in-law, Dr. Nitha Gopalan, who helped me to make my dream come true. I would also like to acknowledge my dear friends for their priceless support and encouragement.

Last but not least, I sincerely thank all those who supported me to complete the thesis in time

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statistical data

Statistics has captured great importance in every field where decisions needed such as agriculture, business, health, technology, law, security and demography. Statistical methods are capable of gathering conclusions from collected data without argue. Ordinarily, data are the reliable information documented for the purpose of analysis. Statistical data are often classified in terms of the number of aspects being studied at a time into; univariate, bivariate and multivariate. The univariate data consists of observations measured only on one attribute. The data which contains a concurrent measurement of two variables referred to as bivariate data, in which it is possible to find the relation between variables. The simultaneous measurement of more than two characteristics of an individual object inherently generates multivariate or multidimensional data and it demands the use of multivariate statistical analysis.

1.2 Multivariate Statistical Methods

Multivariate analysis is a body of methods that are used to obtain statistical inferences from two or more simultaneous measurements from one or more sam-

ples. The measurements refer to variables or individuals and objects refer to units (research units, sampling units, or experimental units) or observations. Historically, immense of applications of multivariate techniques can be seen in the behavioural and biological sciences. However, interest in multivariate methods has now spread to numerous other fields of investigation such as education, chemistry, physics, geology, engineering, law, business, literature, religion, public broadcasting, nursing, mining, linguistics, biology, psychology and so on.

Usually, the variables are measured simultaneously on each sampling unit and these variables are often correlated. If not, there would be little use for many of the techniques of multivariate analysis. It is necessary to untangle the overlapping information provided by correlated variables and peer beneath the surface to see the underlying structure. Thus the goal of many multivariate approaches is simplification. It also aims to express what is going on in terms of a reduced set of dimensions. Such multivariate techniques are exploratory; they essentially generate hypothesis rather than examination. On the other hand, if our goal is a formal hypothesis test, it needs a technique that will (1) allow several variables to be tested and still preserve the significance level and (2) do this for any intercorrelation structure of the variables. A lot of such tests are available in the literature.

In the descriptive realm, one can often obtain optimal linear combinations of variables. The optimality criterion varies from one technique to another, depending on the goal in each case. Although linear combinations may seem too simple to reveal the underlying structure they are often used due to two obvious reasons: (1) they have mathematical tractability (linear approximations are used throughout all science for the same reason) and (2) they often perform well in practice. These linear functions may also be useful as a follow-up to inferential procedures. When statistically significant test results that compare several groups, one can find the linear combination (or combinations) of variables

that leads to the rejection of the hypothesis. Then the contribution of each variable to these linear combinations is of interest. In the inferential area, many multivariate techniques are extensions of univariate procedures. The multivariate inference is especially useful in curbing the researcher's natural tendency to read too much into the data. Some of the widely using multivariate techniques are described below.

1.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a popular statistical method which tries to describe the covariance structure of the data by means of a small number of components. These components are linear combinations of the original variables and often they allow for an interpretation and better understanding of the different sources of variation. In the classical approach, the first component corresponds to the direction in which the projected observations have the largest variance. The second component is then orthogonal to the first and again maximizes the variance of the projected data points. Continuing in this manner produces all the principal components which correspond to the eigenvectors of the empirical covariance matrix.

1.2.2 Factor Analysis

Factor analysis explains the interdependencies among variables in terms of a linear combination of less number of unobservable factors and additional sources of variation. Clearly, factor analysis is a dimension reduction technique that accounts the correlation between observed variables. The contribution of the factors to the variance of an actual variable is termed as factor loadings. The initial interest of factor analysis is the estimation of factor loadings and two popular methods exists for this purpose. At first, the principal component method in which loadings are specified based on the spectral decomposition of the sam-

ple covariance matrix of the data. Another convention of maximum likelihood method used Maximum Likelihood Estimators (MLE) of factor loading with the assumption of normal distribution.

1.2.3 Discriminant Analysis

The objective of discriminant analysis is to separate the objects into mutually exclusive classes on the basis of some apriori information. The primary goal of discriminant analysis is the derivation of a procedure for the optimum allocation. A good classification rule should have a minimum average cost of misclassification. The discrimination can be done by linear composites where each composite is a linear combination of variables. To study the differences among groups, linear combinations of predictor variables are formed. These linear combinations are used to identify the class of an object. The unequal population covariances lead to the quadratic rule where quadratic functions of variables are used for classification. The purpose of the analysis is either to describe group differences or to predict group membership on the basis of response variable measures. The prediction or identification of group membership is done on the basis of one or more predictor or explanatory variables along with one criterion variable. The criterion variable is categorical in nature and measured on a nominal scale. Sometimes it is dichotomous and sometimes it is polytomous.

1.2.4 Canonical Correlation Analysis

Canonical correlation analysis focuses on the maximum correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. It finds a new coordinate system in the space of each set of the variable in such a way that the new coordinates display unambiguously the system of correlation. These linear combinations are the first coordinates in the new system. Then a second linear combination in each set is sought

so that their correlation between them is the maximum of correlations between such linear combinations that are uncorrelated with the first linear combinations. The procedure is continued until the two new coordinate systems are completely specified. The pair of linear combinations is called the canonical variables and their correlations are called canonical correlation.

1.2.5 Multivariate Regression Analysis

Multivariate regression is a methodology that models the relationship between more than one independent(predictors) variables and more than one dependent(responses) variables. It measures the effect of changes in a set of variables corresponding to the changes in another set of independent variables. Similar to classical regression, the main objective of the analysis is to estimate the coefficients of the model. Generally, the coefficient matrices are estimated individually for each dependent variable using the MLE method in multivariate linear regression models. After the model fits the data, the normality of the residual vectors is tested to examine the model adequacy.

1.2.6 Multivariate Analysis of Variance

One-way Multivariate Analysis of Variance (MANOVA) deals with testing the null hypothesis of equal mean vectors across the g considered groups. The setup is similar to that of the one-way univariate Analysis of Variance (ANOVA) but the inter-correlations of the independent variables are taken into account. Under the classical assumptions that all groups arise from multivariate normal distributions, many test statistics are discussed in the literature, one of the most widely used being the likelihood ratio test. This test statistic is better known as Wilks' Lambda in MANOVA. The Wilks' Lambda is reported as part of the test output in almost all statistical packages.

1.2.7 Other Multivariate Techniques

Some tentative methods are accessible to explain the complexity of multivariate data in the form of a measure of distances. Clustering is an important technique of analyzing multivariate data by grouping observations on the basis of similarity among them. Graphical procedures that display multivariate data with different objectivity are multidimensional scaling and correspondence analysis. In first method, distances between observations of multivariate data are visualized into a low-dimensional space whereas correspondence analysis represents the associations in a contingency table. Biplots are the graphical display of multivariate data similar to scatter plots.

Multivariate methods rely on the predetermined model assumptions about the data and these assumptions may not hold in reality due to the presence of outlying observations. Standard computation of many of these multivariate techniques is based on the classical estimation of mean vector and covariance matrix which are excessively influenced by the outlying observations in the data set.

1.3 Outliers

The "outlying" observations have always been a concern in statistical data analysis. Data points that are unrepresentative of the population can mislead the analytical results. These rogue observations are either noise to the data of population or they are an unexpected situation in the natural variability of the population. Clearly, outliers are data units that creates inconvenience in modeling true characteristics of the rest of the dataset. Inconsistency made by outlying observation can influence statistical inferences about the data. Hence the development of appropriate methodology for dealing with exceptional data points that are unnoticed by the traditional analysis is necessary. Some typical definitions

for outliers in literature are given as:

An outlier is defined as "one that appears to deviate markedly from other members of the sample in which it occurs." (Grubbs 1969)

An outlier is defined as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism." (Hawkins 1980)

An outlier is defined as "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." (Barnett and Lewis 1994)

Barnett and Lewis (1994) categorized the causes of outliers as follows:

- *Inherent variability*: this is the natural variability of the population under study.
- *Measurement error*: this consists flaw of measuring equipment or false recording of values.
- *Execution error*: This contains the circumstances including selection of observations which are not in the populations of interest or selection of biased or misjudged samples.

The erroneous observation due to measurement error and execution error can be identified from the data. However, the point due to inherent variability should continue to exist in the data.

Outlier affects the analysis of a dataset in different ways. In 1949 a case of Hadlum vs. Hadlum held in England is an example of identifying an outlier which itself is important. Mr.Hadlum appealed the refusal of an earlier petition for the divorce. The petition is on the ground of Mrs.Hadlum's claimed adultery by giving birth to a child on August 12, 1945, 349 days after Mr.Hadlum had left the country. The gestation period of 349 days is a large outlying observation compared to the average gestation period of 280 days (Barnett and Lewis 1994).

Here the petitioner's interest is to identify outliers as relevant in its own right. Another example in which outlying observations themselves have significance is in predicting terrorist activities explained by Wasi et al. (2014). Fraudulent health care insurance claims identified using matrices designed for outlier detection are discussed in Van Capelleveen et al. (2016). In this case also the flagged outliers are prime important and it is referred for further investigation.

On the other hand, assume a scientist studying a particular type of butterflies. If the data collected contains another type of butterflies with different features, the scientist does not want these outlier observations to influence the statistical inferences of the actual population. The outlying observations that contaminate the distribution of the sample are needed to be identified. The technique which accommodates unrepresentative part of the population, but remain undetected while estimating and analysis are termed as robust. The contaminated distribution model for observed data was established by Tukey (1960). It can be represented as:

$$F = (1 - \alpha)F_0 + \alpha G \tag{1.1}$$

where F_0 is the parametric distribution of genuine observations (clean data) and G is the unknown distribution of contamination and $\alpha < 0.5$ is the proportion of outliers in the data. This is the case where a fraction of rows of a data table may be contaminated. Many methods are available in the literature for robust estimation. Different measures for testing the robustness of the robust estimations methods are reviewed in the following section.

1.4 Robust Estimation of Location Vector and Scatter Matrix in Multivariate Data

Outliers in data on a single variable are simple to define, which is significantly large or small with respect to others. An observation that contains a measurement of more than single variable is termed as a multivariate observation. De-

tection of outliers in multivariate data could be more difficult than univariate methods. In multivariate data, an outlier may not be an extreme point but it can be anywhere in the data. Simple visual identification of multivariate outliers is not feasible as the outliers are not endpoints (Gnanadesikan and Kettenring 1972). Barnett and Lewis (1994), has extensively discussed the significance of outliers in multivariate data and methods to detect the anomaly. Anomalous observations in a single variable can be detected as outliers when a univariate outlier detection method is applied in each variable. But a multivariate observation identified as an outlier refers to a measurement of a combination of the different variable which are significantly deviate.

An important measure of the robustness of an estimator against outlying observations is the breakdown point. The definition of breakdown point is the fraction of arbitrary contaminating observations that can be presented in the sample before the estimate can become arbitrarily large as provided by Hampel (1968) and Hampel (1971), motivated from the discussion of Hodge (1967). More conventional definitions for breakdown value of location vector and covariance matrix have been presented by Lopuhaä and Rousseeuw (1991). For a multivariate random variable $\mathbf{X}$, the breakdown point $\varepsilon_n(\hat{\boldsymbol{\mu}}, \mathbf{X})$ of its location vector $\hat{\mu}$ is defined by:

$$\varepsilon_n(\hat{\boldsymbol{\mu}}, \mathbf{X}) = \underset{m}{min} \left\{ \frac{m}{n}; \underset{\tilde{\mathbf{X}}}{sup} \|\hat{\boldsymbol{\mu}}(\tilde{\mathbf{X}}) - \hat{\boldsymbol{\mu}}(\mathbf{X})\| = \infty \right\} \tag{1.2}$$

where $\tilde{\mathbf{X}}$ is the set of observations contaminated by replacing arbitrary values. On the other hand (1.2) states that the breakdown point of a location vector is the smallest fraction of observations that can be contaminated by outliers before the distance between true sample mean and the distorted sample means can become arbitrarily large.

In the same way the formal definition of the breakdown point of covariance

estimator $\hat{\mathbf{\Sigma}}$ is defined by:

$$\varepsilon_n(\hat{\mathbf{\Sigma}}, \mathbf{X}) = \underset{m}{min} \left\{ \frac{m}{n}; \underset{\tilde{\mathbf{X}}}{sup}\ D\left(\hat{\mathbf{\Sigma}}(\tilde{\mathbf{X}}) - \hat{\mathbf{\Sigma}}(\mathbf{X})\right) = \infty \right\} \tag{1.3}$$

where $D(\mathbf{A}, \mathbf{B}) = max(|\lambda_l(\mathbf{A}) - \lambda_l(\mathbf{B})|, |\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}|)$ and $\lambda_i(\mathbf{A})$ is the i^{th} ordered eigenvalue of $\mathbf{A}$. Expression (1.2) states that breakdown point of covariance estimator is the smallest portion of observations that can be contaminated by the outlying observations before the difference between largest eigenvalues of actual covariance estimate and that of defected covariance estimate becomes arbitrarily large or the difference between the smallest eigenvalues of two estimates is near to zero. It is important to note that the highest limit for the breakdown point is 50%. The breakdown point of the classical estimator of the mean vector and the covariance matrix is $1/N$, where N is the sample size (Donoho and Huber 1982) *i.e* contamination of single observation that may contort the estimate by (1.2) and (1.3). Hence, Rousseeuw and Leroy (1987) presents that in estimating location vector and covariance matrix it is advantageous to use estimator with maximum breakdown point.

Influence function is a significant measure to evaluate the robustness of an estimator. Influence function was introduced by Hampel (1968, 1974), in order to study the infinitesimal behavior of a robust estimator. It was described by Hampel et al. (1986), as the standardized effect of an outlier at the point x on the estimator. It describes the infinitesimal stability of an estimator. Ideally, influence function of a robust estimator should be bounded.

Affine equivariance is another property of a robust estimator. This property guarantees that the estimate will behave in a deterministic way if the samples were subjected to an affine transformation. Specifically, a location estimator $\hat{\boldsymbol{\mu}}$ and covariance estimator $\hat{\mathbf{\Sigma}}$ of a multivariate random variable $\mathbf{X} \in \mathbf{R}^p$ is affine equivariant if and only if, for any vector $\boldsymbol{a} \in \mathbf{R}^p$ and nonsingular square matrix

$\boldsymbol{B}$ of order p.

$$\hat{\boldsymbol{\mu}}(\boldsymbol{B}\mathbf{X} + \boldsymbol{a}) = \boldsymbol{B}\ \hat{\boldsymbol{\mu}}(\mathbf{X}) + \boldsymbol{a} \tag{1.4}$$

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{B}\mathbf{X} + \boldsymbol{a}) = \boldsymbol{B}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{B} \tag{1.5}$$

More clearly, affine equivariant estimates can be converted accordingly for data rotations and location-scale changes. The flexibility to this requirement may increase the number of available robust estimates or may exist in the case of non-affine equivariant estimator that performs better, but affine equivariance is also advantageous to the robust multivariate estimators of location and scatter.

Besides, the effect of masking is also asserting to the breakdown of the outlier detection method in finding multivariate outliers. The inability to detect outliers due to their very presence is termed as masking effect (Wilcox 2017). For example, a small group of outliers of the same direction could vary the mean and increase the standard deviation in such a way that it is undetectable. The amount of masking is measured on the basis of false positives or type II error because of its condition of incorrect decision making of the true outlier. The study of the masking breakdown point was conducted by Becker and Gather (1999). According to Becker and Gather (1999), the masking breakdown value of outlier detection is bounded by the breakdown value of the mean and covariance estimator of that method and it equals the breakdown value of the estimators if both breakdown values are equal. These explanations prove that a single outlier can cause masking of non-robust Mahalanobis distance outlier detection.

Another issue regarding the incorrect identification of possible outliers is termed as swamping effect. This is the phenomenon of identifying uncontaminated observation as outliers. Hadi (1992) referred to the swamping effect as the condition in which all observations with large distances are necessarily not outliers. For instance, a small cluster of outliers will attract the mean and inflate the standard deviation in its direction and away from other observations which belong to the pattern suggested by the majority of observations. One of the

solutions to avoid such false alarms is the use of robust estimates of mean vector and covariance matrix in multivariate outlier detection.

Several methods of distinct perspective have been suggested over many years to detect multiple outliers from multivariate data. As per the conducted survey, the methods can be classified into three: distance-based methods, projection-based methods and other robust methods. In distance-based methods, some form of robust estimators of mean vector and the covariance matrix is defined and then Mahalanobis distance is determined for observations based on these estimates in order to identify the potential outliers whose distance exceeds the specific margin called *threshold value* or *cutoff value*. Then, the mean vector and covariance matrix are robustly estimated using the remaining data points. The projection-based method constitutes the idea that an outlying multivariate observation can be projected into a univariate outlier. The other robust methods take advantage of the statistic of different kinds for manifesting the outlying observations and robust estimation. Generally, these approaches are computationally easier in outlier identification.

1.5 Different Multivariate Robust Methods

Several robust distance-based methods of multivariate outlier detection and estimation of the location vector and scatter matrix have been proposed over the years. Review of these methods are explained below.

1.5.1 Distance-Based Methods

1) Method of M-estimation

M-estimation is one of the earliest distance-based robust estimation procedure of location and scatter introduced by Maronna (1976). Initially, it was proposed by Huber (1964), for the estimation of a univariate location parameter. Campbell (1980), applied this affine equivariant estimation method for outlier detection and

principal component analysis. Later on, Huber (1981), defined M-estimates of mean vector $\mathbf{t}$ and covariance matrix $\mathbf{V}$ as the optimum solutions to the following simultaneous equations:

$$\frac{1}{n}\sum_{1=1}^{n} u_1\left[\left\{(\mathbf{x}_i-\mathbf{t})^T\mathbf{V}^{-1}(\mathbf{x}_i-\mathbf{t})\right\}^{1/2}\right](\mathbf{x}_i-\mathbf{t})=0 \tag{1.6}$$

$$\frac{1}{n}\sum_{1=1}^{n} u_2\left[\left\{(\mathbf{x}_i-\mathbf{t})^T\mathbf{V}^{-1}(\mathbf{x}_i-\mathbf{t})\right\}^{1/2}\right](\mathbf{x}_i-\mathbf{t})(\mathbf{x}_i-\mathbf{t})^T=\mathbf{V} \tag{1.7}$$

where u_1 and u_2 are the weight function of Mahalanobis distance defined on the basis some conditions. The weight functions are basically used for down weighting the effect of outliers in the estimation. Here, the optimality of iterative solutions of equation (1.6) and (1.7) is doubtful. According to Maronna (1976), the major limitation of M-estimator is its least breakdown value $1/(p+1)$ (p is the number of variables) which makes it unsuitable for large dimensional data.

2) MVE and MCD Methods

A high breakdown point alternative for M-estimator, Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) of robust estimation of location and scatter was introduced by Rousseeuw (1985). MVE is based on the computation of the smallest ellipsoid containing at least $h=[n/2]+1$ of the observations of the data, where n is the number of samples. Here, the location vector is the center of the ellipsoid and covariance is the ellipsoid itself. Similarly, MCD searches for the smallest covariance determinant which encompasses at least half of the data points. Mahalanobis distances of each observation are calculated for identifying outliers in both robust methods. The MVE and MCD methods are suitable for highly contaminated situations because of its breakdown value 50%. Complexity in finding optimum subgroup is the major demerit of these robust methods.

Rousseeuw and Leroy (1987), described a resampling algorithm based on the

idea of searching for a small number of good points rather than for bad points in which mean and covariance subset of size $p+1$ drawn from the data are calculated. The corresponding ellipsoid is inflated or deflated to obtain h observations and are repeated to retain minimum volume ellipsoid. Rousseeuw and Leroy (1987), suggested a reweighting technique in order to increase the efficiency of MVE. For reweighed MVE, mean vector and covariance estimates are recalculated only for the set of samples whose Mahalanobis distance corresponding to the initial MVE mean vector and covariance matrix is lower than the suitable appropriate threshold (quantile of Chi-square distribution with p degrees of freedom). Rousseeuw and Van Zomeren (1990), also recommended this type of reweighting in MVE method Lopuhaä and Rousseeuw (1991), proved that the breakdown value of MVE mean vector and covariance matrix are preserved in one step reweighting.

3) Method of S-estimation

In the framework of multiple regression Rousseeuw and Yohai (1984), proposed S-estimator as solution which minimizes the symmetric and continuously differentiable function of residuals. S-estimator shares the properties of M-estimator (Maronna 1976). Then Lopuhaä (1989), extended definition of S-estimator of mean vector and covariance matrix based on the solution ($\mathbf{t}$, $\mathbf{V}$), where $\mathbf{V}$ being a positive definite symmetric matrix, that minimizes $|\mathbf{V}|$ subject to:

$$\frac{1}{n}\sum_{1=1}^{n}\nu\left[\left\{(\mathbf{x}_i-\mathbf{t})^T\mathbf{V}^{-1}(\mathbf{x}_i-\mathbf{t})\right\}^{1/2}\right]=b_0 \tag{1.8}$$

Generally, to ensure the consistency at normal distribution, the constant b_0 can be calculated as $E_{0,I}(\nu \parallel \mathbf{X_0} \parallel)$. The ν-function should be symmetric around zero, twice continuously differentiable and strictly increasing on $[0,k]$ and constant on $[k,+\infty[$. For $\nu(d) = I(|d| > \sqrt{\chi^2_{p,0.5}})$ and $b_0 = 0.5$, provides MVE estimator. S-estimators do satisfy the first order conditions of M-estimators with high breakdown point. Different choices of ν -functions were discussed in

Rocke (1996), which proves that S-estimators are influenced by the presence of outliers even if it has the breakdown value 50%.

4) Hadi's Forward search method

Three major drawbacks of the MVE approach (Rousseeuw and Leroy 1987) are observed by Hadi (1992). First, a number of subsamples must be decided by the user. This suggestion is not feasible because it directly depends on the number of outliers in the data, which is clearly unknown. The second issue is that the covariance estimate for the computation of distance is based on the sub-sample of size $p+1$, which can be unrealistic in practice. The third problem is that the several sub-samples may have singular covariance estimates and different sub-samples may have different covariance. If these are assigned to the user, the MVE estimate will be different as well. Thus, the uniqueness of the MVE estimate is unclear.

A non-affine equivariant MVE based method for multivariate outlier detection is proposed by Hadi (1992), to rectify the limitations of the classical MVE resampling method. For this method, a coordinate-wise median is estimated from the original data and this median vector is used for the estimation of covariance. The observations corresponding to [(n+p+1)/2] smallest distances are selected and classical mean vector and covariance matrix estimates from this sub-group are used to compute Mahalanobis distances for all observations of original data. Again, a subset of size $p+1$ with the smallest distance is chosen and this is being refereed as the basic subset. The sub-sample in the MVE resampling method and the basic subset are different in two aspects. First, the basic subset consists of observations nearest to the centroid and nearness is based on the robust Mahalanobis computed from the coordinate-wise median. Secondly, Hadi's method consists of one basic subset, but resampling MVE methods includes hundreds of

sub-samples and this difference makes Hadi's method computationally simple.

5) Atkinson's Forward search method

Similar to Hadi's forward search method, Atkinson (1993), proposed a forward search algorithm based on the concept of the MVE resampling method. Atkinson's forward search algorithm begins with the estimation of the mean vector and covariance matrix from the randomly selected subset of size $m = p + 1$. Then, the covariance to include h observations of the original data are adjusted and the volume of the covariance is computed. The resulting covariance is then used for the computation of the Mahalanobis distance for all observations. The process is repeated using $m + 1$ observations with smallest distance and during the time, observation whose squared Mahalanobis distance which exceeds the critical value is identified as a potential outlier. Once $m = n$, the process is repeated with a new random subset of size m. Likewise, the algorithm can be executed through the required number of random starting subset. The adjusted covariance matrix that gave minimum volume among all the trials can be used for estimating mean vector, covariance matrix and outlier detection. Atkinson (1993), uses this method in stalactite plots to analyze which of the observation consistently emerged as an outlier. Atkinson (1994), gave a detailed illustration of this method as well.

6) Hawkin's Feasible solution algorithm

To reduce the exhaustive enumeration, Hawkins (1994), suggested a Feasible Solution Algorithm (FSA) for obtaining global optimum for MCD estimator (Rousseeuw 1985). FSA procedure starts with an assumption that there exists atmost k outlier in the data. Sample of $n - k$ observations are randomly selected from the original data to form an initial subset and the remaining k observations from the data are trimmed. The mean vector and covariance matrix are esti-

mated from this random subset. Study each pair of observations for which one observation from the random subset and replace it with one from the remaining trimmed set. Compute the covariance determinant of the new subset and compare it with the determinant of old covariance estimate. If the new determinant is smaller than the old one, the old covariance estimate is useless and update it with the new one. This process is repeated for all pairs and find the subset of $n - k$ observations that produces greatest reduction in covariance determinant. The whole process can be repeated with other random subsets for an additional feasible solutions. The subset which has minimum determinant is then considered for obtaining final MCD estimates.

7) Hybrid algorithm

The method based on robust distances has been discussed in the literature for a number of strategies, that include: 1) combinatorial methods, in particular, MVE and MCD; 2) smooth estimators such as M-estimator, referred by Rocke and Woodruff (1996); and 3) forward search algorithms proposed by Hadi (1992) and Atkinson (1993). In order to combine the FSA by Hawkins (1994) and forward search algorithms, Rocke and Woodruff (1996), suggested a hybrid algorithm. It is also an exploration of Rocke (1996), Rocke and Woodruff (1993), Woodruff and Rocke (1993) and Woodruff and Rocke (1994). This affine equivariant multivariate outlier detection method consists of two phases. The purpose of Phase I is to obtain robust estimates of location and shape for the dataset and this problem fall in the combinatorial and smooth methods. At first, Hawkins's FSA is used to obtain an approximate MCD estimate of location and shape. The resulting estimates are substituted as the starting point of Atkinson's forward search method against estimates of mean vector and covariance matrix of a random subset of $p + 1$ point originally suggested by Atkinson. Then, a set of observations that are free of outlier identified by the Atkinson's method is used

to obtain initial estimates of mean vector and covariance matrix. This heuristics share the property of the better result of forward search method with a good starting point and the globally optimum M-estimation is close to this solution. Basically, combinatorial methods such as MCD seeks space that increases exponentially with sample size. The data is partitioned to user-defined cells to cope with these defects. The robust estimator of mean vector and covariance matrix are then obtained with minimum covariance determinant.

The results of Phase I are used for computing Mahalanobis distance to all observations of Phase II. Scale the resulting distances in order to be consistent with the distances obtained from multivariate normal data and compare the scaled distances to a suitable critical value (suitable quantile of Chi-square distribution with p degrees of freedom).

8) Method of Resampling by Half-Mean and Smallest Half-Volume

The multivariate outlier detection techniques such as MCD and M-estimator rely on large computational effort and restrict the usefulness in high-dimension. Egan and Morgan (1998), introduced two simple methods to detect outliers in high-dimensional data: 1) Resampling by Half-Mean (RHM) method 2) Smallest Half-Volume (SHV) method. In the RHM method, $n/2$ observations are randomly selected from the dataset without replacement and each observation is used to make a new matrix, i^{th} sample matrix $\mathbf{X}_{(i)}$. Then, mean and standard deviation of each column of $\mathbf{X}_{(i)}$ are computed and used for rescaling the original data matrix. The magnitude of each observation of rescaled observations is calculated, which is equal to the distance of each rescaled observations from the centroid of data.

The computed distances are used to form i^{th} column of distance matrix $\mathbf{L}$. After the desired number of samples are generated, each column of matrix $\mathbf{L}$ is sort in ascending order. The largest 5% of values in each column of $\mathbf{L}$ are

identified, observations that appear in the largest 5% with high frequency are detected as outliers. The main drawback of this method is the unclear idea and subjective decision about the number of appearances that indicate outliers.

The second intuitive outlier detection method, SHV method begins by standardizing each column of original data using the respective column mean, which is termed as auto-scaling. The Euclidean distance between each observation is then computed and used to form a $n \times n$ matrix with zero diagonal elements. Each column of the distance matrix is sorted in ascending order and the sum of the first $n/2$ observations is measured. The column with the lowest sum is identified and the observations that are used to measure the sum are characterized as good observations. Then the subset with good observations is used to estimate mean vector and covariance matrix and used as robust estimates for classical Mahalanobis distance to detect outliers.

9) Bivariate Box Plot Method

A less-formal method, box plot for detecting univariate outlier reveal the location, spread, and skewness of the data graphically. To construct a bivariate box plot and then to detect multivariate outliers, Zani et al. (1998), proposed a method in which the inner region is determined through convex hull peeling (Bebbington 1978). Convex hull peeling technique begins by identifying and trimming the observations on the convex hull of the data and repeating this process until a specific percentage of original data remains. Bivariate plot proposed by Zani et al. (1998), suggested to trim the data until 50% of observations remain in the set. The remaining observations create the inner region of bivariate box plot. The method of B-splines (Ammeraal 1992) make sure the smoothness of contours of the inner region. The arithmetic mean of observations remaining in the inner region defines the centroid of the bivariate box plot. Zani et al. (1998), propounded to create a bivariate box plot for each pair of variables. Remove the

observations which fall outside 90% convex hull in any box plot, 75% is appropriate for small sample sizes. Then, the rest of the observations are used for the starting point of the forward search method by Hadi (1992, 1994) or Atkinson (1994). The assertion behind this method is that the initial subset for the forward search should contain more than $p + 1$ observations, which makes further analysis of outliers computationally effective.

10) Partitioning Method

After computational experiments Rocke and Woodruff (1996), observed that the proposed hybrid algorithm is inadequate in detecting outliers from contaminated multivariate data when fraction of outliers is 35% or more. Besides, Kosinski (1998), recognized the obtained half-sample based on MCD does include outliers. To reduce this insufficiency of outlier identification at highly contaminated situations Kosinski (1998) proposed an alternative outlier detection method that aims to find a partition of data that distinguishes good observations from outlying ones. This partitioning method is a repetition of Hadi's and Atkinson's forward search method in a way that is applied to get multiple random starting subset. The choices of initial subsets are specified to ensure that at least one good partition contains none of the outliers.

11) FAST-MCD Method

The MVE and MCD method are initially proposed by Rousseeuw (1985). Among these method MVE gained more consideration in multivariate outlier detection on the grounds of simplicity in computation. Later, Butler et al. (1978) has shown that MCD is asymptotically normal and this leads to better statistical efficiency than MVE and Davies (1992) proved that MVE has lower convergence rate. Based on the observations of Rousseeuw and Van Driessen (1999), MCD has better theoretical advantages than MVE. However, MCD lacks computational in-

tricacies to find the half-sample subset with a minimum covariance determinant. To minimize the drawback, Rousseeuw and Van Driessen (1999), constructed an algorithm for MCD which is claimed to be faster than MVE algorithms. The initial theorem by Rousseeuw and Van Driessen (1999), states that one can order all observations in terms of Mahalanobis distance based on the mean vector and covariance matrix of an initial random subset from the original data. Using these estimates, compute the Mahalanobis distance and select a new subset with smallest distance. The covariance determinant of the new subset will be less than or equal to covariance determinant of the old subset. This process referred by the theorem is termed as C-step. Applying C-step several times to the dataset converges to an optimum MCD solution. The examination based on the experiments indicates that C-step is used only two times for the convergence of the solution. Further, ten different subsets with minimum determinant are used for the third subset for C-step convergence.

12) BACON Method

According to Billor et al. (2000), better multivariate outlier detection methods may not reduce the computational complexity of robust estimation. Also observed that the affine-equivariant estimators may add substantial computational complexity to the methods without a proportional improvement in outlier identification. To attain computationally efficient robust estimator and multivariate outliers detection methods, Billor et al. (2000), proposed the Blocked Adaptive Computationally Efficient Outlier Nominator (BACON) method that uses iterative procedure without optimality conditions. BACON methods includes two versions of outlier detection algorithms, first works as an affine-equivariant with breakdown point of 20% and the second works as an approximately affine-equivariant with breakdown point of 40%. The BACON method is derived from Hadi's forward search method (Hadi 1992, 1994) and the algorithm begins by

selecting an initial outlier free subset. The initial basic subset can be chosen in two ways, in the first case the initial basic subset consists of $p+1$ observations with smallest Mahalanobis distance based on the estimates of mean vector and covariance matrix of entire dataset. In the second case, the initial subset consists of $p+1$ observations with the smallest Mahalanobis distance based on the component-wise median vector and covariance matrix derived from the component-wise median vector. The second algorithm gives a more robust and less affine equivariant estimator since the component-wise median is not affine-equivariant. The mean vector and covariance matrix of the initial basic subset are then used to compute the Mahalanobis distance of entire observations. These computed distances are compared to a suitable quantile of Chi-Square distribution with p degrees of freedom.

13) OGK Method

On the contrary to the early methods of time complexity, Maronna and Zamar (2002), proposed an Orthogonalized Gnanadesikan-Kettenring (OGK) estimator in which a general method is applied to make a scatter matrix estimate, positive-definite and approximately affine-equivariant. The non positive-definite matrix is constructed by applying the robust covariance estimator introduced by Gnanadesikan and Kettenring (1972), to each pair of variables. The resulting scatter matrix and component-wise medians are then orthogonalized to make robust estimates of location vector and covariance matrix for outliers detection. This approximately affine-equivariant method performs better even under highly collinear situations. The re-weighting steps are applied to improve the efficiency of the OGK estimators.

14) MCD-EHD Method

Generally, univariate outlier detection methods use iterative deletion tech-

nique in which most distant observations are deleted from the data. Then, delete the second most observation, repeat the iteration until no other observations are identified. Caroni and Prescott (1992) introduced a sequential approach to the Wilks (1963) test for a single outlying observation from a multidimensional sample. Further, Viljoen and Venter (2002) proposed Minimum Covariance Determinant-Extreme Hotelling Deviate(MCD-EHD) method that used the sequential procedure with a starting subset based on the FAST-MCD algorithm proposed by Rousseeuw and Van Driessen (1999).

15) RMCD Method

The limitations of outlier detection methods that rely on the comparison of the robust distances with the Chi-Square distribution are pointed out by Cerioli (2010). The Chi-Square approximation of distances to decide the critical value may be adequate if it is certain that the data contains outliers. Caroni and Prescott (1992) developed a sequence of critical values to make simultaneous corrections in comparing distances of the entire data. To obtain good performance in the robust estimation, even though the data is free of outliers, Cerioli (2010) proposed a Reweighted Minimum Covariance Determinant (RMCD) method. This procedure begins with the MCD estimate of mean vector and covariance matrix, multiplied by a correction factor to attain consistency and unbiasedness on the multivariate normal distribution. The squared robust Mahalanobis distances are calculated on the basis of obtained MCD estimates and weight 0 is assigned to corresponding observation whose distance exceeds the threshold value (suitable quantile of Chi-Square distribution with p degrees of freedom) to get subset with smallest distances. The robust estimates mean vector and covariance matrix computed using these subsets are called as RMCD estimates, the correction factor guarantees consistency of RMCD covariance estimate. Cerioli (2010), also proposed an iterated version of RMCD, on the intention of increasing the power

of multiple outlier detection rules.

16) Optimized Method

A more optimized outlier detection method to compete with MVE and MCD methods, specifically for data that contains fewer outliers was developed by Oyeyemi and Ipinyomi (2010). Initially, all possible combinations of a subset of size $p+1$, where p is the dimension of the dataset, are obtained. Oyeyemi and Ipinyomi (2010), were interested to select the sub-sample such that eigenvalues of the estimated covariance matrix satisfy three optimality criteria; it should be the minimum of minimum eigenvalues, minimum of the product of eigenvalues, minimum of the harmonic mean of eigenvalues. Utilizing this sub-sample, the mean vector and covariance matrix are estimated to compute Mahalanobis distance and select $p+1$ least distant observations. Repeat this process until the sub-sample include $h = (n+p+1)/2$ observations, where, n is the size of the original sample and p is the dimension. However, optimized robust method for Hotelling's T^2 control outperformed MVE and MCD for very limited sample sizes and dimensions.

17) Comedian Method

In spite of the computational complexity, robust methods such as MVE, MCD, and OGK are affected by swamping and masking problems. In the interest of reducing the insufficiency, Sajesh and Srinivasan (2012) developed a robust procedure for estimating the mean vector and covariance matrix, known as comedian method. Falk (1997) introduced a non-positive semi-definite robust covariance estimator known as the comedian that generalizes Median Absolute Deviation(MAD) (Huber 1981). The comedian method of outlier detection utilizes the properties of the highest breakdown point along with equivariance of comedian and correlation median and a robust correlation estimate is derived

from it. To solve the lack of positive definiteness and affine-equivariance of co-median estimator, the orthogonalization technique provided by Maronna and Zamar (2002), was applied. The Mahalanobis distance is calculated for all entire data based on the orthogonalized component-wise median and comedian matrix. The observations corresponding to the distance that exceeds an adjusted critical value are identified as outliers. The rest of the observations are used to subsequent robust estimation of mean vector and covariance matrix. Sajesh and Srinivasan (2012), recommended a reweighing process to increase the efficiency of outlier detection. This method is stated to be more efficient in high dimensional data with less masking and swamping effect.

18) DetMCD Method

The Majority of robust estimators of the mean vector and covariance matrix include the computational burden of drawing a large number of sub-samples to at least one initial set of uncontaminated observations. In the case of the MCD estimator, introduced by Rousseeuw (1985), its computation would not be feasible till FAST-MCD is established (Rousseeuw and Van Driessen 1999). Hubert et al. (2012), proposed a deterministic way of approximation of the MCD denoted as DetMCD, possibly faster than FAST-MCD.

The selection of sub-samples in DetMCD would not be random but the iterations are the same as that of FAST-MCD. To make the process, location and scale equivariant, each variable is standardized by subtracting its median and dividing by Q_n estimate proposed by Rousseeuw and Croux (1993). The primal estimates of the mean vector and scatter matrix are then subjected to the eigenvector transformations. Then, these estimates are substituted for the calculations of Mahalanobis distances of all observations. Thus, the robust distances were calculated using different orthogonalized estimates, observations of size $\lceil n/2 \rceil$ with least distance were selected to estimate robust distances again

and applied C-steps up to convergence.

DetMCD method used different scatter estimators as initial estimator which include classical correlation matrix, Spearman rank correlation matrix, correlation of normalized quantiles, spatial sign covariance, mean vector and covariance matrix produced by the first algorithm of BACON method (Billor et al. 2000) and OGK estimates in which, median and Q_n were the initial location and scale estimates. Among these estimates, the result corresponding to the minimum covariance determinant is chosen as raw DetMCD estimates, re-weighting steps are employed for final DetMCD estimates. The DetMCD algorithm neither depends on a single initial estimate nor a unique objective, the algorithm confides on the six estimators of a different perspective. Deterministic nature as well as the time complexity related to the randomness in the C-step and duration of the initial estimates.

19) DetS and DetMM Methods

Fast-S estimation method was developed by Salibián-Barrera, Van Aelst and Willems (2006) to minimize the computation complexity of S-estimator (Lopuhaä 1989), in optimizing the objective function. For Fast-S estimator, scatter matrix $\mathbf{V}$ is replaced by shape matrix $\sigma^2\mathbf{\Gamma}$ in equation (1.8), where σ be the scale estimate, $\mathbf{\Gamma} = |\mathbf{V}|^{-1/p}\mathbf{V}$ be the shape matrix selected such that $|\mathbf{\Gamma}| = 1$ and p be the dimension of $\mathbf{V}$, so that $|\mathbf{\Gamma}| = 1$ always. Therefore, the objective is to estimate the triplet $(\hat{\mathbf{t}}, \hat{\mathbf{\Gamma}}, \hat{\sigma})$ that minimizes s such that,

$$\frac{1}{n}\sum_{1=1}^{n} \nu \left[\frac{\left\{(\mathbf{x}_i - \mathbf{t})^T \mathbf{\Gamma}^{-1} (\mathbf{x}_i - \mathbf{t})\right\}^{1/2}}{s} \right] = b_0 \tag{1.9}$$

where $\mathbf{t} \in \mathbf{R}^p$, $\mathbf{\Gamma}$ be semi-positive definite square matrix of order p, $|\mathbf{\Gamma}| = 1$, and s be a positive scalar.

The fast algorithm starts with selecting N sub-samples at random from the original dataset. Sub-group solutions using equation (1.9) are denoted as,

$(\hat{\mathbf{t}}_l^0, \hat{\mathbf{\Gamma}}_l^0, \hat{\sigma}_l^0)$, for $l = 1, ..., N$, where $\hat{\sigma}_l^0$ is median of Mahalanobis distances using $\hat{\mathbf{t}}_l^0$ and $\hat{\mathbf{\Gamma}}_l^0$. Then, these estimates are improved by applying I-step, The j^{th} I-step refinement of scale estimate begins with,

$$\hat{\sigma}_l^j = \hat{\sigma}_l^{j-1} \frac{1}{nb_0} \left[\sum_{1=1}^{n} \nu \left[\frac{\left\{ (\mathbf{x}_i - \hat{\mathbf{t}}_l^{j-1})^T (\hat{\mathbf{\Gamma}}_l^{j-1})^{-1} (\mathbf{x}_i - \hat{\mathbf{t}}_l^{j-1}) \right\}^{1/2}}{\hat{\sigma}_l^{j-1}} \right] \right]^{1/2}$$

Thus, the weighted mean $\hat{\mathbf{t}}_l^j$ and weighted covariance $\hat{\mathbf{V}}_l^j$ $\hat{\sigma}_l^j$ determined using weights $W_i^j = \nu'(u)/u$, where $u = \dfrac{\left\{ (\mathbf{x}_i - \hat{\mathbf{t}}_l^{j-1})^T (\hat{\mathbf{\Gamma}}_l^{j-1})^{-1} (\mathbf{x}_i - \hat{\mathbf{t}}_l^{j-1}) \right\}^{1/2}}{\hat{\sigma}_l^{j-1}}$ and results refinement of $\hat{\mathbf{\Gamma}}_l^j = |\hat{\mathbf{V}}_l^j|^{-1/p} \hat{\mathbf{V}}_l^j$. The scale estimates can be improved by applying the refinement until convergence by keeping mean and scatter fixed. Like wise refine the triplet estimates for specific number of typically less than l smallest iterated scales and not for all initial sub-group estimates. Choose mean vector and covariance $\hat{\mathbf{V}}^F = (\hat{\sigma}^F)^2 \hat{\mathbf{\Gamma}}^F$ corresponding to the smallest scale among all and after complete refinements as final estimates.

Motivated by the high breakdown point and robustness of DetMCD method, Hubert et al. (2015) proposed DetS estimator to increase the robustness of Fast-S estimation. DetS method begins by calculating six primary scatter estimates of scales variables using three-step orthogonalization in OGK method much like DetMCD. The shape matrix and scale are estimated using each estimate. Further, These values replace the initial triplet estimates to six sub-samples in Fast-S estimator denoted by $(\hat{\mathbf{t}}_l^0, \hat{\mathbf{\Gamma}}_l^0, \hat{s}_l^0)$, for $l = 1, ..., 6$, . Then, continue the remaining processes of k I-step in the Fast-S estimation using scaled variables and determined six estimates. The scale estimates are refined again by fixing estimates of mean vector and shape matrix. Select two smallest scale estimates among them and perform I-step until convergence. The resulting estimates $(\hat{\mathbf{t}}^F, \hat{\mathbf{\Gamma}}^F, \hat{s}^F)$ are used to estimate location vector and scatter matrix of original data *i.e*, $\hat{\mathbf{t}}(\mathbf{X}) = \mathbf{Q}\hat{\mathbf{t}}^F + median(\mathbf{X})$ and $\mathbf{V}(\mathbf{X}) = \mathbf{Q}\hat{\mathbf{V}}^F\mathbf{Q}$, where

$\mathbf{Q} = diag(Q_n(X_1), ..., Q_n(X_p)$ and $\hat{\mathbf{V}}^F = (\hat{s}^F)^2\hat{\mathbf{\Gamma}}^F$.

Along with DetS Hubert et al. (2015) proposed DeMM estimator. The procedure of DetMM estimation consists of two steps. At first, the multivariate estimates of mean vector and covariance matrix $(\tilde{\mathbf{t}}, \tilde{\mathbf{V}})$ are obtained by solving objective function of S-estimation in equation (1.8) and the estimate $\tilde{s} = |\tilde{\mathbf{V}}|^{1/2p}$. Secondly, find $(\hat{\mathbf{t}}, \hat{\mathbf{\Gamma}})$ which minimizes

$$\frac{1}{n}\sum_{1=1}^{n} u_1 \left[\frac{\left\{(\mathbf{x}_i - \mathbf{t})^T \mathbf{\Gamma}^{-1} (\mathbf{x}_i - \mathbf{t})\right\}^{1/2}}{\tilde{s}} \right]$$

where $\mathbf{t} \in \mathbf{R}^p$ and $\mathbf{\Gamma}$ is a semi-positive definite matrix with $|\mathbf{\Gamma}| = 1$. This yields the DetMM estimator of location $\hat{\mathbf{t}}$ and scatter $\hat{\mathbf{V}} = (\hat{s})^2\hat{\mathbf{\Gamma}}$.

Hubert et al. (2015) found that DetS method performs better than Fast-S method that is an arbitrary number of subsamples to ensure at least one clean subset, which is unrealistic. Also observed that both robust methods are permutation invariant and almost affine-equivariant.

20) MDP Algorithm

The traditional methods of robust distance are unreliable when the dimension of data exceeds sample size. To address the problems with high dimensionality, a modified Mahalanobis distance was suggested by Ro et al. (2015), defined to be:

$$d_i^2(\boldsymbol{\mu}, \mathbf{D}) = (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{D} (\mathbf{x}_i - \boldsymbol{\mu}) \tag{1.10}$$

where, $\mathbf{D}$ diagonal matrix with diagonal elements of covariance matrix $\mathbf{\Sigma}$. Ro et al. (2015), proposed the Minimum Diagonal Product (MDP) algorithm, which includes two sets of algorithms using modified distance. Similar to the MCD method, the first algorithm aims to find a subset of h observations in which product of marginal variances will be minimum and the estimates $\hat{\boldsymbol{\mu}}_{MDP}$, $\hat{\mathbf{\Sigma}}_{MDP}$ and $\hat{\mathbf{D}}_{MDP}$ are accordingly the sample mean, sample covariance and diagonal

matrix with diagonals of covariance matrix. If multiple solutions occur in this minimization, MDP estimators are selected arbitrarily. Repeat the process by choosing the second set of h observations with the smallest distances in equation (1.10) by applying previous MDP estimates. This procedure is employed for m initial subsets. Next, choose the subset with MDP value for further estimation. Ro et al. (2015), used reweighting, presented by Cerioli (2010), to increase the efficiency of MDP.

The second algorithm computes initial estimates of first algorithm for $h = [n/2]+1$ where $\hat{\mathbf{D}}_{MDP}$ is multiplied with a consistency constant. Finally, second step is reweighted using the threshold value, a function of estimated correlation matrix. Ro et al. (2015), suggested that MDP algorithm is fast even when $n = 100$ and $p = 400$.

21) MRCD Method

MCD estimators are not applicable when dimension exceeds sample size. Hence to confront this situation Boudt et al. (2019) proposed, Minimum Regularized Covariance Determinant (MRCD) method. For a multivariate data $\mathbf{X}$, of order $n \times p$, the MRCD method which generalizes the MCD method was introduced by Rousseeuw (1985), begins with standardized variable like in DetMCD method. Boudt et al. (2019), then regularized covariance matrix for target matrix ($\mathbf{T}$) and regularization parameter (ρ) is defined as $\mathbf{K}(\mathbf{H}) = \rho\mathbf{T}+(1-\rho)c_\alpha\mathbf{S_U}(\mathbf{H})$, where $\mathbf{S_U}(\mathbf{H})$ is the covariance estimator using MCD method for standardized data $\mathbf{U}$ and c_α is the consistency factor.

Singular value decomposition is then performed for $\mathbf{T}$, *i.e* $\mathbf{T} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ and estimated transformed standardized observations $w_i = \mathbf{\Lambda}^{-1/2}\mathbf{Q}^T\mathbf{u}_i$. It follows that $\mathbf{S_W}(\mathbf{H}) = \mathbf{\Lambda}^{-1/2}\mathbf{Q}^T\mathbf{S_U}(\mathbf{H})\mathbf{Q}\mathbf{\Lambda}^{-1/2}$. To find MRCD subsets, six robust and well-conditioned initial estimates of location $\mathbf{m}^i$ and scatter $\mathbf{S}^i$ are obtained by using different initial scatter estimator of Hubert et al. (2012). Then, select the subsets

of $\mathbf{W}$ consisting h observations with lowest Mahalanobis distance based on each $\mathbf{m}^i$ and $\mathbf{S}^i$. The smallest value ρ^i is then determined for each well-conditioned regularized covariance. Then set $\rho = \underset{i}{max}\ \rho^i$ if $\rho^i \leq 0.1$, $\rho = max\{0.1, \underset{i}{median}\ \rho^i\}$. If $\rho^i \leq \rho$ for initial subset, repeat the C-step using the standardized and regularized covariance of the last subset and ensure that the final MRCD subset has the lowest determinant of regularized covariance. The final estimates of the mean vector and covariances are estimated using the obtained subset. This method aims to regularize the covariance before minimizing its determinant and claimed to be performed better and faster than MCD in all ranges of dimensions.

22) SM Method

Considering median is a robust measure of central tendency, multivariate outlier detection methods are constructed based on component wise median. Multivariate expansion of univariate median is termed as spatial median (Brown 1983). Applying spatial median as the robust measure of multivariate location and covariance based on Sajana and Sajesh (2018b) developed Spatial Median (SM) method based on Mahalanobis distance. The SM method performed efficiently with low masking and swamping for at most 25% contamination in the data.

23) Other Distance Based Methods

Parsimony issues in the low sample and high dimension was discussed in Ahn, Lee and Zi (2018) and presented a robust distance-based solution to handle this problem. To address similar situations in high dimensional data, Yang et al. (2018), proposed threshold-based method for outlier detection. Leys et al. (2018) studied the outliers of different viewpoints of experimental social psychology. They suggested a variant Mahalanobis distance-based method to address this issue. Weighted likelihood estimators of multivariate location and scatter

have been presented by Agostinelli and Greco (2019). The proposed methods performed adequately for a small proportion of outliers.

1.5.2 Projection-based Methods

1) Stahel-Donoho estimator

Stahel-Donoho estimator was independently proposed by Stahel (1981) and Donoho (1982), to estimate the mean vector and covariance matrix in which observations are down-weighted relative to some projection of data to one-dimensional space. It is found that the estimators asymptotically attain a breakdown value of 50%. Rousseeuw and Leroy (1987), applied Stahel-Donoho estimator in the calculation of robust distance. Stahel-Donoho measure of *outlyingness* of a multivariate observation $\mathbf{x}_i$ for a p-dimensional projection vector $\mathbf{v}$ is defined as:

$$u_i = \sup_{\|\mathbf{v}\|=1} \frac{\mid \mathbf{x}_i\mathbf{v}^T - \underset{j}{med}(\mathbf{x}_j\mathbf{v}^T) \mid}{\underset{k}{med} \mid \mathbf{x}_k\mathbf{v}^T - \underset{j}{med}(\mathbf{x}_j\mathbf{v}^T) \mid} \tag{1.11}$$

Further, mean vector and covariance matrix are estimated using measured u_i of each observation as follows:

$$T(\mathbf{X}) = \frac{\sum_{i=1}^{n} w(u_i)\mathbf{x}_i}{\sum_{i=1}^{n} w(u_i)} \tag{1.12}$$

$$V(\mathbf{X}) = \frac{\sum_{i=1}^{n} w(u_i)(\mathbf{x}_i - T(\mathbf{X}))(\mathbf{x}_i - T(\mathbf{X}))^T}{\sum_{i=1}^{n} w(u_i)} \tag{1.13}$$

where $w(u_i)$ is a decreasing weight function. Stahel-Donoho method is able to combine affine equivariance and breakdown property together. But the computational complexity in the calculation of outlyingness of the observations essentially adverse the properties and limit its practical use in outlier detection. However, Gasko and Donoho (1982), provided a robustification using these estimators for

leverage diagnostics in multiple regression.

2) Projection Pursuit Detection

To refrain from methods that used Mahalanobis distance in identifying outliers and to avoid the masking and swamping effect, Pan et al. (2000), proposed a projection pursuit method. This method projects multivariate data into univariate observations, then used the univariate method to detect outliers from the projected observations. Thus, the outlier identifier creates a Gaussian process on p-dimensional unit hypersphere. Pan et al. (2000), recommended that the method flags relatively less false positives. On the other hand, there is no specific indication of the number of projection points needed to scatter on the unit hypersphere to generate multivariate data. Therefore the application in large dimensional dataset will be difficult.

3) Kurtosis Method

The computational difficulties of projection-based methods such as the Stahel-Donoho algorithm rely on the amount of randomly generated directions. Reduction of the number of successful directions is a solution to improve the efficiency of these kinds of methods. Peña and Prieto (2001) suggested a method in which kurtosis coefficients are used to obtain directions. The sample points are projected on to a set of $2p$ directions, where p is the dimension of the data. These directions are determined through minimizing and maximizing the kurtosis coefficients of projected points.

The kurtosis coefficient measures the peakedness of a distribution. Outliers from the symmetric contamination model increase the kurtosis coefficient. A small amount of asymmetrical outliers leads to a higher kurtosis coefficient as well. Conversely, it decreases to a very small value for a large number of asymmetrical outliers. Accordingly, obtaining the directions that minimize and maximize

the kurtosis coefficient of projection would be advantageous to outlier detection (Peña and Prieto 2001).

Determination of direction is a global optimization problem, which is not practically efficient instead local minimizers and local maximizers are computed. Peña and Prieto (2001) showed that these local optimizers assure to find either: direction to the outlier or a direction orthogonal to it. The uncertainty of obtained directions carries off the procedure of projecting the data onto a subspace orthogonal to the computed directions and other directions obtained. Then, this process is repeated to attain p direction for the maximizer of the projected kurtosis coefficient and p direction for the minimizer of projected kurtosis coefficient. To decide the outlyingness of observation on any of these $2p$ directions based on the univariate median and Median Absolute Deviation (MAD), the maximum distance of observation form the median exceeds a suitable cutoff value. Thus, mean and covariance estimates are calculated using observations that are not outliers. These estimates are then used to compute Mahalanobis distance for entire data and the observations are labeled as outliers whose distance exceeds the desired quantile of χ^2 distribution with p degrees of freedom. The performance of kurtosis coefficient directions was not satisfactory for large contamination levels. To improve the result of this situation Peña and Prieto (2007), suggested a method based on the combination of specific direction and random directions. Stratified sampling is applied to generate random directions. This semi-deterministic procedure was an improvement of the Kurtosis method.

1.5.3 Other Robust Methods

Most of the distance-based methods constraint the choice of at least one uncontaminated initial subset. The iterative computations to select an outlier-free subsample becomes time consuming. An alternative identification principle other than distance has equal importance in the robust field. To detect the lack of fit

of individual observation, Rao (1964) suggested a method that uses a sum of squared length of projections other than robust distance. This technique has been applied by Gnanadesikan and Kettenring (1972), to detect outliers. To determine the deviated observations from the linear space of the smallest principal components, the sum of squares is calculated. The large value of sum indicates the outlyingness of the corresponding observation. The reliability of these methods entrusted on the subjectivity in identifying outliers. A procedure based on the decomposition of Mahalanobis distance was explained by Kim (2000). On the assumption of normality, the Mahalanobis distance is decomposed and provides a component scatter plot for placing the outliers.

An outlier detection method based on the estimated angles has been proposed by Juan and Prieto (2001) to separate the clustered outliers. They described that the angle between distribution (uniform distribution) of uncontaminated data and the reference direction (u_0) is a function of the Beta distribution. After obtaining the angles, they suggested to drawing the Q-Q plot to test presence of outliers using the lack of fit. The spacing test for goodness-of-fit has been studied by Pyke (1965). Then, the distribution of spacing between each consecutive ordered observations is determined. The cutoff of outlier identification has been calculated from the spacing distribution using the largest interval of normalized spacing. Chiang et al. (2003) proposed an algorithm that makes use of PCA as well as significant tests. To test whether the observation is an outlier, the $p - a$ the eigenvectors are used, where p is the dimension of original data. Q-statistic, presented by Jackson and Mudholkar (1979), used to test the outlyingness of the observations. If the Q-statistic of an observation exceeds the respective threshold specified by Chiang et al. (2003), then the observations are flagged as outliers and repeat the process flagged between iterations.

Another subjective procedure based on eigenvalue and eigenvector was established by Gao et al. (2005), referred to as Max-Eigen Difference (MED) method.

The initial step of this method is the determination of eigenvalue and eigenvector of the scatter matrix of the original data. The covariance matrix is calculated for each dataset with i^{th} observation removed and eigenvalue and eigenvector are obtained for these covariance matrices. As specified by Gao et al. (2005), the observation with a larger MED value is marked as outliers. Kirschstein et al. (2013) proposed pruning Minimum Spanning Tree (pMST) method that considers the entire dataset in a network. The MST method begins by designing a sphere around each observation with a radius. The number of spheres in the largest connected set of spheres is used to find the subsample of good observations and can be used for robust estimation purposes. Obviously, the observations that are not part of the largest connected sphere are identified as outliers. When there are more a fraction of outliers and fraction of good samples, the pMST method performs better than MCD. But for other cases of higher dimensions, the MCD is better.

An algorithm is proposed by Liebscher et al. (2013), in which the Delaunay triangulation is adopted to select the uncontaminated subset. Wang and Zwilling (2015) established an outlier detection based on the Voronoi diagram explained in Preparata and Shamos (1985). It is automatically constructing the neighborhood relationship of the observations in original data.

1.6 Aims and Objectives of the Study

The literature review included traditional and non-traditional methods, some of them are derived and determined uniquely for the purpose of robust estimation and others are the compound of earlier strategies. The discussion of previous studies has shown that all methods have commendable properties as well as limitations. The main limitation, however, is the effect due to masking and swamping. Most of these methods undergo highly complicated computations as well. Another challenge lies in the breakdown point of the outlier detection and

robust estimation. The maximum proportion of outliers that the estimate can safely tolerate is termed as the breakdown value of an estimator. Similarly, the maximum proportion of outliers that can successfully detect can be defined as the breakdown value of an outlier detection method. Many of these methods fail when the data contains more than 20% outliers. There are a few methods with a higher level of breakdown points, but still, there is scope for improvement.

In this thesis, an attempt is made to propose a multivariate outlier detection procedure that considers the primary issues confronted in the preceding study. It also presents a robust estimator for mean vector and covariance matrix based on the outlier detection method which can deal with multiple outliers from large dimensional data. The performance evaluation is conducted in the classical contamination model in terms of Rate of Successful Detection (RSD) and Rate of False Detection (RFD). RSD measures the rate of successful detection of true outliers while RFD represents the false detection of inliers as outliers. In fact, RSD and RFD are the way of assessing the masking effect and swamping effect respectively. The RSDs and RFDs can also be used to measure the breakdown value. Moreover, the efficiency of the robust estimates of the mean vector and the covariance matrix can be evaluated using Mean Squared Errors (MSE). Robust statistical techniques based on the proposed method have been discussed and evaluated.

1.7 Outline of the Thesis

This thesis is organized as six chapters including the current chapter on the introduction of outliers, robust estimation and is organized as follows.

Chapter 2, consists of a distance-based outlier detection method for a robust estimation of multivariate linear regression coefficients. The efficiencies of the proposed estimates compared with some popular methods and results are provided in terms of mean square errors.

Chapter 3, presents a robust alternative for covariance estimator based on robust scale estimator by repeated median measurements. On the basis of proposed robust bivariate dispersion, an robust correlation estimator is proposed and the efficiencies are examined by calculating weighted mean square errors.

Chapter 4, provides a method to detect outlying observation from multidimensional data and subsequent robust estimation of mean vector and covariance matrix. A distance-based approach and adjusted threshold are adopted to identify the outliers in the multivariate dataset. The performance of the proposed method is evaluated using masking and swamping effect of the method in Monte Carlo experiments.

Chapter 5, presents an application of the proposed robust estimation method in discriminant analysis which is one of the significant multivariate statistical technique for data analysis. The produced robust discriminant analysis is examined using simulated training dataset and validation dataset. The overall misclassification probabilities are computed for different classification datasets of various ranges of contaminations.

Chapter 6, contains summary and discussion of the main results of the thesis has been presented along with scope and directions for future research.

Chapter 2

Robust Estimation of Multivariate Linear Regression Parameters[1]

2.1 Introduction

The regression analysis has an important part in studying the relationship between variables in statistical data. It consists of modeling, analysis of several variables and aims to find the relationship between response variables and predictor variables. Clearly regression analysis focuses on understanding the changes made by differences of predictor variables in response variable. In the multivariate regression model, the study is between multiple predictor variables and multiple response variables. Thus the multivariate regression model is defined as follows

$$\mathbf{y}_i = \mathbf{B}\mathbf{x}_i + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, i = 1, ...n \tag{2.1}$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{ip}) \in \mathbf{R}^p$, $\mathbf{y}_i = (y_{i1}, ..., y_{iq}) \in \mathbf{R}^q$. The $p \times q$ slope matrix $\mathbf{B}$, the intercept vector $\boldsymbol{\alpha} \in \mathbf{R}^q$ and independently and identically distributed random variables $\boldsymbol{\varepsilon}_i$ with zero mean and symmetric and positive definite scatter $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ are the parameters in multivariate linear regression model (Rousseeuw et al. 2004).

The location vector $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$ of the joint variable $(\mathbf{x}, \mathbf{y})$ can

[1]Some part of this chapter is based on Sajana and Sajesh (2018a)

be partition in the following form,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}$$

Let MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be denotes as $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. Therefore MLE of multivariate regression parameters in equation (2.1) are defined as

$$\hat{\mathbf{B}} = \hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\Sigma}}_{xy} \tag{2.2}$$

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}_y - \hat{\mathbf{B}}^T\hat{\boldsymbol{\mu}}_x \tag{2.3}$$

$$\hat{\boldsymbol{\Sigma}}_{\varepsilon} = \hat{\boldsymbol{\Sigma}}_{yy} - \hat{\mathbf{B}}^T\hat{\boldsymbol{\Sigma}}_{xx}\hat{\mathbf{B}} \tag{2.4}$$

These estimators are inappropriate for the situations where the data contains outliers since MLE is highly affected by the presence of outliers. Outliers are observations inconsistent with the remainder of the dataset (Barnett and Lewis 1994). Hence, classical estimation methods are not robust against presents of outliers in the dataset (Barnett and Lewis 1994). As a solution to this problem, one may replace these classical estimates by highly robust estimates which are less sensitive to outliers and perform robust analysis.

Koenker and Portnoy (1990) investigated the application of M-estimator to each coordinate of the responses and Bai et al. (1990) suggested to minimize the sum of the Euclidean norm of residuals for estimating multivariate linear regression models, these methods are not affine equivariant. An overview of the robust multivariate regression technique is explained by Maronna and Yohai (1997) in the context of simultaneous equation models. Ollila et al. (2002) introduced the robust estimation procedure of multivariate regression based on the sign covariance matrix. Rousseeuw (1985) discussed the application of MCD method which possesses a high breakdown point and Rousseeuw et al. (2004) developed the reweighted versions of the MCD estimator in multivariate regression estimation. The MCD methods provide better robust estimates of multivariate regression

coefficients compared to many other methods, the computational complexity involved in this method limits its application. Sajesh and Srinivasan (2013) studied different techniques for outlier detection in multidimensional data.

Sajesh and Srinivasan (2012) proposed a highly robust technique called *comedian* method for the estimation of mean vector and covariance matrix. They showed that comedian estimates posses high breakdown point. The method of *comedian* approach was adopted by Sajana and Sajesh (2018a) for the estimation of multivariate linear regression parameters. The efficiency of the proposed method is evaluated through empirical experiments and the results are compare with MLE, MCD and OGK estimators. The performance the method at real scenario is illustrated using bench mark datasets.

2.2 Robust Estimation of Multivariate Regression Coefficients

Let $\mathbf{Z}$ be a $n \times (p+q)$ data matrix consisting of p predictor variables and q response variables. Consider $\mathbf{z}_i$ $(i = 1, 2, .., n)$ and $\mathbf{Z}_j$ $(j = 1, 2, .., (p+q))$ are the rows and columns of the data matrix. Thus, the *comedian* matrix $\mathbf{COM}(\mathbf{Z})$ is defined as

$$\mathbf{COM}(\mathbf{Z}) = (COM(\mathbf{Z}_i, \mathbf{Z}_j)), i, j = 1, 2, ..., p+q \tag{2.5}$$

where $COM(\mathbf{Z}_i, \mathbf{Z}_j) = med\{(\mathbf{Z}_i - med(\mathbf{Z}_i))(\mathbf{Z}_j - med(\mathbf{Z}_j))\}$. Since $\mathbf{COM}(\mathbf{Z})$ is not positive semi-definite (Falk 1997), to solve this issue the orthogonalization technique to attain positive-definite and approximately scatter matrices, described by Maronna and Zamar (2002) is applied. In order to do so a robust correlation matrix is then determined based on $\mathbf{COM}(\mathbf{Z})$ is defined as

$$\boldsymbol{\delta}(\mathbf{Z}) = \mathbf{DCOM}(\mathbf{Z})\mathbf{D}^T \tag{2.6}$$

where $\mathbf{D}$ is the diagonal matrix with elements $1/MAD(\mathbf{Z}_i)$ $(1 = 1, 2, ..., p+q)$ and $MAD(\mathbf{Z}_i) = med(|\mathbf{Z}_i - med(\mathbf{Z}_i)|)$

The following orthogonalization process is used to obtain the robust estimates of location vector and scatter matrix for estimating multivariate linear regression parameters. Consider a square matrix $\mathbf{E}$ with columns are eigenvectors of $\boldsymbol{\delta}(\mathbf{Z})$. Let $\mathbf{Q} = \mathbf{D}(\mathbf{Z})^{-1}\mathbf{E}$ and $\mathbf{w}_i = \mathbf{Q}^{-1}\mathbf{z}_i$, $i = 1, 2, ..., n$. Thus $\mathbf{W}$ be the orthogonalized matrix with rows $\mathbf{w_i}$ $(i = 1, 2, ..., n)$ and columns $\mathbf{W}_j$ $(j = 1, 2, ..., (p + q))$. The resulting robust estimates of location vector and scatter matrix are then defined as

$$\boldsymbol{\mu}_R = \mathbf{Ql} \; and \; \boldsymbol{\Sigma} = \mathbf{Q\Gamma Q}^T \tag{2.7}$$

where $\mathbf{l} = (med(\mathbf{W}_1), ..., med(\mathbf{W}_{p+q}))$ and $\boldsymbol{\Gamma} = diag(MAD(\mathbf{W}_1)^2, ..., MAD(\mathbf{W}_{p+q})^2)$. The procedure can be iterated to compute $\boldsymbol{\Sigma}_R$ and $\boldsymbol{\mu}_R$ for $\mathbf{W}$, then expressing them in the original coordinate system. These estimates can be improved by a reweighting step using the robust Mahalanobis distance defined as,

$$rd_i = (\mathbf{z}_i - \boldsymbol{\mu}_R)^T \boldsymbol{\Sigma}_R^{-1} (\mathbf{z}_i - \boldsymbol{\mu}_R), \; i = 1, 2, ..., n \tag{2.8}$$

Let m be the weighting function, then define weighted mean vector and covariance matrix respectively as,

$$\boldsymbol{\mu}_{RW} = \frac{\sum_i^n m_i \mathbf{z}_i}{\sum_i^n m_i} \; and \; \boldsymbol{\Sigma}_{RW} = \frac{\sum_i^n m_i (\mathbf{z}_i - \boldsymbol{\mu}_{RW})(\mathbf{z}_i - \boldsymbol{\mu}_{RW})^T}{\sum_i^n m_i} \tag{2.9}$$

where weight function $m_i = 1$ for $rd_i \leq cv$ and equals 0 otherwise and the cutoff value (cv) of hard rejection is determined as

$$cv = 1.4826 \frac{\chi^2_{p+q,0.95} \; med(rd_1, ..., rd_{n)}}{\chi^2_{p+q,0.5}} \tag{2.10}$$

Thus, the resulting estimates of equation (2.9) can be partitioned to attain the comedian estimates of multivariate linear regression parameter and are defined by

$$\hat{\mathbf{B}}_R = \hat{\boldsymbol{\Sigma}}_{Rxx}^{-1} \hat{\boldsymbol{\Sigma}}_{Rxy} \tag{2.11}$$

$$\hat{\boldsymbol{\alpha}}_R = \hat{\boldsymbol{\mu}}_{Ry} - \hat{\mathbf{B}}_R^T \hat{\boldsymbol{\mu}}_{Rx} \tag{2.12}$$

$$\hat{\boldsymbol{\Sigma}}_{R\varepsilon} = \hat{\boldsymbol{\Sigma}}_{Ryy} - \hat{\mathbf{B}}_R^T \hat{\boldsymbol{\Sigma}}_{Rxx} \hat{\mathbf{B}}_R \tag{2.13}$$

2.3 Finite Sample Efficiency

To investigate the finite sample efficiency of *comedian* multivariate regression, following simulation study is performed. Generated m datasets of sample size (n) from a distribution $N_{p+q}(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. The experiment has been conducted for different choices of n, p and q to ensure the performance evaluation of various situations. For each dataset $\mathbf{Z}^{(k)}$ $(k = 1, 2, ..., m)$, *comedian* method has been carried out for yielding $p \times q$ slope matrix $\widehat{\mathbf{B}}^{(k)}$, intercept vector $\widehat{\boldsymbol{\alpha}}^{(k)}$ and $q \times q$ covariance matrix estimate $\widehat{\boldsymbol{\Sigma}}_{\epsilon}^{(k)}$ of errors. To measure the sample efficiency, MSE is computed. The MSE of a univariate estimator is defined to be

$$MSE(T) = n \underset{k}{ave}(T^{(k)} - \theta)^2$$

where θ is the true value of the parameter. Similarly MSE of $\widehat{\mathbf{B}}$ and $\widehat{\boldsymbol{\alpha}}$ are defined as

$$MSE(\widehat{\mathbf{B}}) = \underset{i,j}{ave}(MSE(\widehat{\mathbf{B}}_{i,j})),\ i = 1, 2, ..., p \ and \ j = 1, 2, ..., q \tag{2.14}$$

$$MSE(\widehat{\boldsymbol{\alpha}}) = \underset{j}{ave}(MSE(\widehat{\boldsymbol{\alpha}}_j)),\ j = 1, 2, ..., q \tag{2.15}$$

Likewise for the diagonal and off-diagonal of $\widehat{\boldsymbol{\Sigma}}_{\epsilon}$

Table 2.1 gives the results of the comparative study. The results of the proposed comedian regression is compared with results of MLE, MCD and OGK regression estimators. All simulation are performed with $m = 1000$ replications for various values of n between 50 and 1000. From the table it is clear that, in uncontaminated situation *comedian* regression has the MSE same as that of MLE estimation method, but this is not true for other methods. Simulations for

Table 2.1: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.202	1.051	1.014	1.002
Intercept	1.176	1.032	1.003	1.004
$\mathbf{\Sigma}_{diagonal}$	2.641	2.126	2.065	2.050
$\mathbf{\Sigma}_{off-diagonal}$	0.897	0.983	0.996	0.997
MLE:				
Slope	1.201	1.051	1.013	1.002
Intercept	1.176	1.032	1.003	1.004
$\mathbf{\Sigma}_{diagonal}$	2.641	2.126	2.065	2.050
$\mathbf{\Sigma}_{off-diagonal}$	0.897	0.983	0.996	0.997
MCD:				
Slope	3.571	1.541	1.245	1.171
Intercept	2.290	1.297	1.137	1.086
$\mathbf{\Sigma}_{diagonal}$	6.323	2.883	2.412	2.325
$\mathbf{\Sigma}_{off-diagonal}$	3.187	1.540	1.235	1.179
OGK:				
Slope	1.732	1.428	1.353	1.329
Intercept	1.522	1.260	1.225	1.205
$\mathbf{\Sigma}_{diagonal}$	5.322	5.774	8.046	12.440
$\mathbf{\Sigma}_{off-diagonal}$	0.908	1.056	1.074	1.093

other sample sizes n and different dimensions p and q gives similar results and are presented in Appendix A.

Outliers in regression can be commonly classified into two, vertical outliers and bad leverage points. Vertical outliers are the observations that do not follow the linear pattern of the majority of the data due to the outlyingness of response variables. On the other hand, leverage points are outlying observations due to the outlyingness of the predictor variables. It can be classified into good leverage points and bad leverage points with regards to the observation that follows or does not follow the pattern of rest of the data.

The datasets are generated with vertical outliers and bad leverage points to study the proposed method in contaminated situations. In order to include vertical outliers in the data, sample of size n is generated from $N_{p+q}(\mathbf{0}, \mathbf{I})$ then replace the $100\gamma\%$ (γ=rate of contamination) observations of each response variable by $N(2\sqrt{\chi^2_{p+q,0.99}}, 0.1)$. The finite sample results from data with $\gamma = 0.1, 0.2$ vertical outliers are shown in Table 2.2 and Table 2.3 respectively. The MSE comparison for other choices of contamination and dimension is presented in Table A.2, Table A.3 and Table A.4 of Appendix. The MSE values form these tables indicated that the *co-median* regression produces less error compared to MLE, MCD and OGK methods.

To incorporate the bad leverage points in the data, sample of size n is generated from $N_{p+q}(\mathbf{0}, \mathbf{I})$ then replace the $100\gamma\%$ observations of p predictor variables and q response variables by independently generated samples from $N(2\sqrt{\chi^2_{p,0.99}}, 0.1)$ and $N(2\sqrt{\chi^2_{q,0.99}}, 0.1)$. The simulation is repeated $m = 1000$ times for computing MSE values of *comedian* regression estimates. The results of experiment conducted for data with bad leverage points are presented in Table 2.4 and Table 2.5 respectively for $\gamma = 0.1, 0.2$. In the same manner, the finite sample error values for $\gamma = 0.4$ and different variable dimensions are provided in Table A.5, Table A.6 and Table A.7 of Appendix. From these tables it is also evident that, MSE of *comedian* regression are less affected by the presence of bad leverage points when compared to MLE, MCD and OGK estimates.

In the previous tables, the data with independent variables are considered. For the purpose of investigating the behavior of *comedian* regression estimator in the correlated dataset, $m = 1000$ datasets of sample size n from a distribution $N_{p+q}(\mathbf{0}, \mathbf{I_{0.5}})$, where $\mathbf{I_{0.5}}$ is the covariance matrix with unit variance and covariance=0.5 is generated. Further, the presence of vertical outliers and bad leverage points are also studied for correlated dataset. The MSE values of *comedian* regression estimator, MLE, MCD and OGK for vertical outliers and bad leverage

Table 2.2: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$ when the data contains 10% of vertical outliers

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.335	1.144	1.141	1.109
Intercept	1.299	1.121	1.150	1.147
$\mathbf{\Sigma}_{diagonal}$	2.935	2.421	2.283	2.226
$\mathbf{\Sigma}_{off-diagonal}$	0.996	1.060	1.129	1.092
MLE:				
Slope	12.545	10.731	10.792	10.537
Intercept	54.249	211.368	523.930	1051.373
$\mathbf{\Sigma}_{diagonal}$	3424.488	16510.274	42681.683	86414.403
$\mathbf{\Sigma}_{off-diagonal}$	3584.411	16937.652	43621.652	88193.685
MCD:				
Slope	3.226	1.516	1.344	1.264
Intercept	2.229	1.344	1.236	1.205
$\mathbf{\Sigma}_{diagonal}$	6.288	4.373	5.875	9.717
$\mathbf{\Sigma}_{off-diagonal}$	3.336	1.749	1.556	1.459
OGK:				
Slope	1.876	1.515	1.515	1.508
Intercept	1.651	1.365	1.349	1.357
$\mathbf{\Sigma}_{diagonal}$	4.988	5.050	6.501	9.192
$\mathbf{\Sigma}_{off-diagonal}$	0.964	1.132	1.226	1.248

Table 2.3: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$ when the data contains 20% of vertical outliers

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.608	1.346	1.355	1.349
Intercept	1.491	1.282	1.303	1.260
$\mathbf{\Sigma}_{diagonal}$	3.501	2.901	2.850	2.932
$\mathbf{\Sigma}_{off-diagonal}$	1.073	1.213	1.239	1.236
MLE:				
Slope	21.442	18.3134	18.259	17.677
Intercept	212.177	839.032	2100.356	4193.410
$\mathbf{\Sigma}_{diagonal}$	10932.928	52231.799	134780.7	272669.8
$\mathbf{\Sigma}_{off-diagonal}$	11340.839	53554.466	137901.5	278789.0
MCD:				
Slope	4.270	1.585	1.437	1.401
Intercept	10.351	1.401	1.350	1.283
$\mathbf{\Sigma}_{diagonal}$	656.347	8.218	16.027	30.803
$\mathbf{\Sigma}_{off-diagonal}$	669.310	2.057	1.865	1.825
OGK:				
Slope	2.121	1.662	1.659	1.696
Intercept	1.799	1.486	1.564	1.591
$\mathbf{\Sigma}_{diagonal}$	5.109	4.558	5.328	7.223
$\mathbf{\Sigma}_{off-diagonal}$	1.022	1.257	1.371	1.488

Table 2.4: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$ when the data contains 10% of bad leverage points

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.369	1.209	1.191	1.179
Intercept	1.372	1.148	1.148	1.124
$\mathbf{\Sigma}_{diagonal}$	2.989	2.514	2.628	2.70
$\mathbf{\Sigma}_{off-diagonal}$	0.968	1.077	1.111	1.140
MLE:				
Slope	2.562	6.396	14.293	27.484
Intercept	1.469	1.385	1.555	1.759
$\mathbf{\Sigma}_{diagonal}$	2.328	2.480	3.458	4.951
$\mathbf{\Sigma}_{off-diagonal}$	1.989	5.311	11.881	22.086
MCD:				
Slope	3.158	1.547	1.329	1.265
Intercept	2.496	1.355	1.238	1.158
$\mathbf{\Sigma}_{diagonal}$	6.544	4.438	6.095	10.023
$\mathbf{\Sigma}_{off-diagonal}$	3.224	1.737	1.511	1.496
OGK:				
Slope	1.857	1.537	1.499	1.511
Intercept	1.671	1.3408	1.341	1.302
$\mathbf{\Sigma}_{diagonal}$	5.013	4.743	6.465	9.535
$\mathbf{\Sigma}_{off-diagonal}$	0.948	1.137	1.207	1.280

points with $\gamma = 0.1$ exhibited in Table 2.6 and Table 2.7. The results in the correlated cases are same as that of uncorrelated cases.

2.4 Robustness Properties of Comedian Regression Estimator

The robustness properties of the proposed estimator is studied in terms of breakdown point and affine equivariance.

2.4.1 Breakdown Point

Sajesh and Srinivasan (2012) proposed an empirical method to find the break-

Table 2.5: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$ when the data contains 20% of bad leverage points

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.609	1.381	1.333	1.309
Intercept	1.559	1.317	1.299	1.290
$\mathbf{\Sigma}_{diagonal}$	3.539	3.012	2.963	2.938
$\mathbf{\Sigma}_{off-diagonal}$	1.089	1.202	1.238	1.284
MLE:				
Slope	2.753	6.640	14.725	28.236
Intercept	1.717	1.583	1.692	1.877
$\mathbf{\Sigma}_{diagonal}$	3.362	3.306	3.754	4.614
$\mathbf{\Sigma}_{off-diagonal}$	1.687	4.456	10.021	19.102
MCD:				
Slope	6.360	9.697	16.943	30.317
Intercept	5.372	4.302	3.340	3.179
$\mathbf{\Sigma}_{diagonal}$	8.228	14.267	18.583	27.013
$\mathbf{\Sigma}_{off-diagonal}$	3.159	5.111	9.340	16.748
OGK:				
Slope	1.945	1.607	1.558	1.609
Intercept	1.836	1.519	1.528	1.608
$\mathbf{\Sigma}_{diagonal}$	5.302	4.907	5.593	7.172
$\mathbf{\Sigma}_{off-diagonal}$	1.032	1.240	1.324	1.414

Table 2.6: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$ when the correlated data contains 10% of vertical outliers

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.372	1.209	1.191	1.120
Intercept	1.324	1.142	1.148	1.169
$\mathbf{\Sigma}_{diagonal}$	3.029	2.444	2.628	2.228
$\mathbf{\Sigma}_{off-diagonal}$	1.509	1.412	1.111	1.375
MLE:				
Slope	12.269	10.638	10.507	10.585
Intercept	55.468	211.066	526.237	1048.54
$\mathbf{\Sigma}_{diagonal}$	3464.352	16519.968	42702.460	86430.679
$\mathbf{\Sigma}_{off-diagonal}$	3540.396	16713.615	43120.248	87225.074
MCD:				
Slope	3.233	1.575	1.322	1.265
Intercept	2.514	1.364	1.229	1.397
$\mathbf{\Sigma}_{diagonal}$	6.501	4.244	5.856	6.423
$\mathbf{\Sigma}_{off-diagonal}$	4.185	2.340	2.555	1.782
OGK:				
Slope	1.913	1.535	1.477	1.470
Intercept	1.622	1.369	1.328	1.397
$\mathbf{\Sigma}_{diagonal}$	4.848	4.498	5.301	6.423
$\mathbf{\Sigma}_{off-diagonal}$	1.909	1.780	1.839	1.782

Table 2.7: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$ when the correlated data contains 10% of bad leverage points

	n			
	50	200	500	1000
Comedian regression:				
Slope	2.369	5.246	11.336	21.539
Intercept	0.743	0.661	0.649	0.657
$\mathbf{\Sigma}_{diagonal}$	13.915	41.994	99.519	194.702
$\mathbf{\Sigma}_{off-diagonal}$	2.491	1.291	2.769	5.208
MLE:				
Slope	2.583	6.499	14.611	28.170
Intercept	0.714	0.662	0.695	0.729
$\mathbf{\Sigma}_{diagonal}$	14.491	46.623	111.574	218.959
$\mathbf{\Sigma}_{off-diagonal}$	0.489	1.385	3.104	5.937
MCD:				
Slope	4.300	5.585	11.469	21.610
Intercept	1.497	0.775	0.683	0.677
$\mathbf{\Sigma}_{diagonal}$	10.776	31.335	75.158	146. 56
$\mathbf{\Sigma}_{off-diagonal}$	1.304	1.729	3.423	6.406
OGK:				
Slope	2.734	5.563	11.729	22.052
Intercept	0.914	0.767	0.761	0.766
$\mathbf{\Sigma}_{diagonal}$	17.139	50.168	117.845	229.100
$\mathbf{\Sigma}_{off-diagonal}$	0.444	1.167	2.456	4.646

Table 2.8: MSE comparison for breakdown point

p, q	Normal data				48% contaminated with vertical outliers			
	Slope	Intercept	$\mathbf{\Sigma}_{diagonal}$	$\mathbf{\Sigma}_{off-diagonal}$	Slope	Intercept	$\mathbf{\Sigma}_{diagonal}$	$\mathbf{\Sigma}_{off-diagonal}$
6,6	1.000	1.008	2.037	1.008	1.000	1.008	2.037	1.008
6,10	1.012	1.016	2.020	0.991	1.012	1.016	2.020	0.991
10,6	1.002	0.983	2.110	0.987	1.002	0.983	2.110	0.987
10,10	1.017	1.022	2.107	0.982	1.017	1.022	2.107	0.982
15,15	1.013	1.005	2.244	0.986	1.013	1.005	2.244	0.986

Table 2.9: MSE comparison for breakdown point

p, q	Normal data				48% contaminated with bad leverage points			
	Slope	Intercept	$\mathbf{\Sigma}_{diagonal}$	$\mathbf{\Sigma}_{off-diagonal}$	Slope	Intercept	$\mathbf{\Sigma}_{diagonal}$	$\mathbf{\Sigma}_{off-diagonal}$
6,6	1.065	1.160	2.456	0.997	1.065	1.160	2.456	0.997
6,10	1.048	1.030	2.332	1.006	1.048	1.029	2.321	1.005
10,6	1.047	1.027	2.544	0.997	1.047	1.026	2.536	0.996
10,10	1.048	1.019	2.416	1.014	1.048	1.019	2.416	1.014
15,15	1.044	1.023	2.472	1.001	1.044	1.023	2.470	1.001

down value of an outlier detection method. Similarly, the breakdown value of regression estimator can be computed using the finite sample MSE. For the purpose of obtaining breakdown value of proposed estimator, a sample of size n is simulated from the distribution $N_{p+q}(\mathbf{0}, \mathbf{I})$ and compute the MSE of estimates. Then insert additional $100\gamma\%$ percentage of contamination into the existing data and measure the MSE values of *comedian* regression estimates of entire dataset. Identify maximum amount of outliers as breakdown point that the estimator can tolerate with out deviating MSE values.

The breakdown point study consists of two kinds of contamination: vertical outliers and bad leverage points as described in the previous section. Empirical comparison of MSE to determine the breakdown value for vertical outliers and bad leverage points are shown in Table 2.8 and Table 2.9 respectively. From these tables, the MSEs for normal (uncontaminated) dataset and contaminated data set seems equal for different combination of dimensions up to $\gamma = 0.48$.

2.4.2 Affine Equivariance

Rousseeuw and Leroy (1987) generalized the affine equivariance properties of multivariate regression estimators. Adding a linear function of the predictor variable to the response variable is equivalent to adding the linear coefficients with the estimator is termed as regression equivariance. Assume that $\mathbf{T}(\mathbf{X}, \mathbf{Y}) = (\widehat{\mathbf{B}}^T, \widehat{\boldsymbol{\alpha}})$, where $\mathbf{X}$ is $n \times p$ matrix and $\mathbf{Y}$ is $n \times q$ matrix. Then, the estimator $\mathbf{T}$ is said to be regression equivariant if

$$\mathbf{T}(\mathbf{X}, \mathbf{Y} + \mathbf{XC} + \mathbf{I_n v^T}) = \mathbf{T}(\mathbf{X}, \mathbf{Y}) + (\mathbf{C^T}, \mathbf{v})^\mathbf{T} \tag{2.16}$$

where $\mathbf{C}$ is any $p \times q$ matrix, $\mathbf{v}$ is any $q \times 1$ vector and $\mathbf{I}_n = (1, ..., 1)^T \in \mathbf{R}^n$

The estimator $\mathbf{T}$ is said to be y-affine equivariant, if the linear transformation of response variables and the corresponding linear transformations and this can be expressed as

$$\mathbf{T}(\mathbf{X}, \mathbf{YM} + \mathbf{I_n r^T}) = \mathbf{T}(\mathbf{X}, \mathbf{Y})\mathbf{M} + (\mathbf{O_{pq}^T}, \mathbf{r})^\mathbf{T} \tag{2.17}$$

where $\mathbf{M}$ be any nonsingular $q \times q$ matrix, $\mathbf{r}$ is any $q \times 1$ vector and $\mathbf{O_{pq}}$ is the $p \times q$ matrix consisting of zeros. Similarly the estimator $\mathbf{T}$ is said to be x-affine equivariant if

$$\mathbf{T}(\mathbf{YN^T} + \mathbf{I_n d^T}, \mathbf{Y}) = (\widehat{\mathbf{B}}^T\mathbf{N}^{-1}, \widehat{\boldsymbol{\alpha}} - \widehat{\mathbf{B}}^T\mathbf{N}^{-1}\mathbf{d})^T \tag{2.18}$$

where $\mathbf{N}$ is any nonsingular $p \times p$ matrix and $\mathbf{d}$ is any $p \times 1$ vector.

The equivariance properties are empirically proven with the help of simulated samples and MSE in all possible situations by varying parameter. Table 2.10 shows the finite sample results of affine equivariance and it contains the MSE values of the transformed data and transformed estimates of untransformed data according to left hand side and right hand sides of the equivariance equations

Table 2.10: MSE comparison for different affine equivariance with $n = 1000$

p, q	Transformed data					
	Regression Equivariance		Y Equivariance		X Equivariance	
	$\mathbf{B}$	$\boldsymbol{\alpha}$	$\mathbf{B}$	$\boldsymbol{\alpha}$	$\mathbf{B}$	$\boldsymbol{\alpha}$
6,6	336.409	332.911	0.389	330.204	0.374	1.839
6,10	333.429	326.539	0.385	336.376	0.378	1.931
10,6	335.287	342.927	0.370	338.949	0.375	2.441
10,10	333.188	340.942	0.379	335.341	0.379	2.465
15,15	334.682	330.545	0.375	332.042	0.379	3.064
p, q	Transformed estimate					
	Regression Equivariance		Y Equivariance		X Equivariance	
	$\mathbf{B}$	$\boldsymbol{\alpha}$	$\mathbf{B}$	$\boldsymbol{\alpha}$	$\mathbf{B}$	$\boldsymbol{\alpha}$
6,6	336.409	332.911	0.389	330.204	0.377	1.841
6,10	333.429	326.539	0.385	336.376	0.379	1.901
10,6	335.163	342.927	0.370	338.949	0.373	2.380
10,10	333.188	340.942	0.379	335.341	0.377	2.444
15,15	334.682	330.545	0.375	332.042	0.379	3.019

(2.16), (2.17) and (2.18). From the table, it is possible to see that MSE values for both cases are equal and this indicates that the proposed comedian possess affine equivariance property.

Another significant advantage of proposed regression estimators is the less time complexity. A simulation is performed different choices of n and this process is repeated 1000 times, the average time for an estimation procedure is tabulated for different methods is given in Table 2.11. The table shows that the *comedian* regression method required relatively less time for estimation than other methods.

Table 2.11: Average time consumption of different method in R Programming

n	p, q	Time(s)		
		Comedian	MCD	OGK
50	4,4	0.017	0.030	0.023
100	4,4	0.021	0.090	0.027
500	4,4	0.030	0.219	0.038
1000	4,4	0.064	0.249	0.086
50	6,6	0.040	0.069	0.059
100	6,6	0.034	0.162	0.052
500	6,6	0.057	0.421	0.082
1000	6,6	0.083	0.340	0.117
50	10,10	0.089	0.159	0.139
100	10,10	0.115	0.467	0.176
500	10,10	0.139	0.982	0.212
1000	10,10	0.209	1.519	0.317

2.5 Illustration Using Example

The Pulp-Fiber dataset presented by Lee (1992) consists of 62 observations from four predictor variables and four response variables. The predictor variable describes properties of pulp fiber: arithmetic fiber length, long fiber fraction, fine fiber fraction and zero span tensile. The response variables measured: breaking length, elastic modulus, stress at failure and burst strength. Rousseeuw et al. (2004) studied the dataset and showed that the observations 46-48 and 58-62 can be considered as outliers

The diagnostic plot of Pulp-Fiber data is presented in Figure 2.1 in which the robust residual distances plotted against robust distances of the observations. The vertical and horizontal cutoff lines are chosen to be $\sqrt{\chi^2_{4,0.975}}$. From the figure, observations 56, 58, 59, 60, 61 and 62 lie far from both the cutoff lines, these six sample points are then identified as outliers (bad leverage points). The observations 22, 28, 51 and 52 are vertical outliers since they have small residuals. Thus the *comedian* regression estimates of the Pulp-Fiber data are,

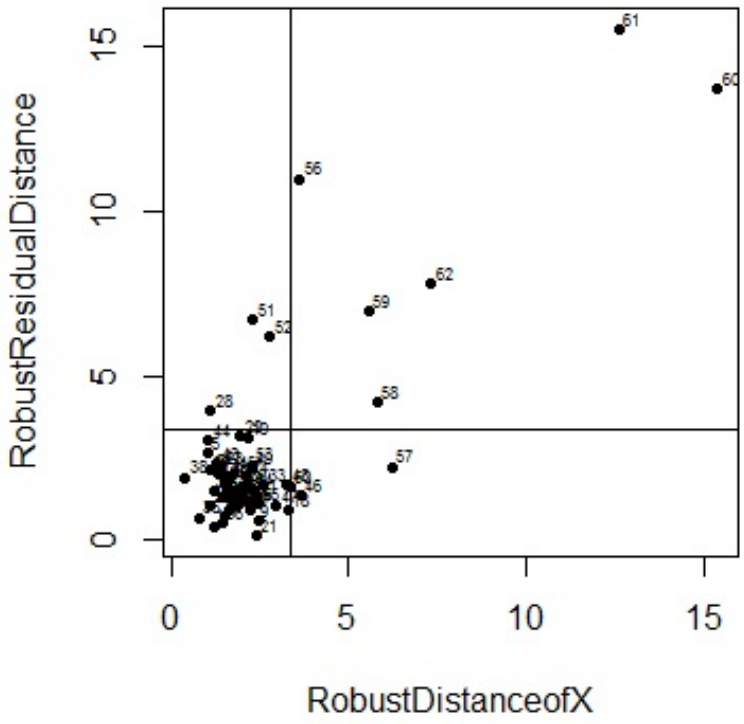


Figure 2.1: Diagnostic plot for Pulp-Fiber data

$$\hat{\mathbf{B}}_R = \begin{pmatrix} -4.814 & -1.727 & -2.515 & -0.719 \\ 0.257 & 0.077 & 0.137 & 0.052 \\ 0.109 & 0.035 & 0.058 & 0.020 \\ 78.448 & 21.292 & 38.523 & 17.117 \end{pmatrix}$$

$$\hat{\boldsymbol{\alpha}}_R^T = \begin{pmatrix} -74.785 & -19.466 & -42.363 & -19.758 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_{R\varepsilon} = \begin{pmatrix} 0.632 & 0.175 & 0.317 & 0.150 \\ 0.175 & 0.058 & 0.088 & 0.043 \\ 0.317 & 0.088 & 0.160 & 0.075 \\ 0.150 & 0.043 & 0.075 & 0.039 \end{pmatrix}$$

2.6 Summary

In this chapter a *comedian* based method is proposed for the robust estimation of multivariate regression parameters. The efficacy of the proposed method is

investigated through simulation study and compared results with MLE, MCD and OGK methods using the parameter MSE. In all the experiments, MSE values of *comedian* regression estimators are much less compared to that of other three methods. The simulation study for time complexity shows that the proposed method is much faster than other methods. A benchmark dataset has been used to investigate the efficiency of the proposed method in real-life situation and it is identified that the proposed method is suitable for the robust estimation of multivariate regression coefficients.

Chapter 3

Robust Estimation using S_n covariance[1]

3.1 Introduction

Bivariate data consists of measurements of exactly two variables and statisticians are always interested to establish the relationship between these variables. Simply, the direction of the linear relationship between random variables can be measured using covariance estimation. Correlation coefficient is a scaled measure of association between two random variables. Unfortunately, classical covariance and correlation coefficients are not robust against the existence of possible outliers. Hence it is necessary to propose a robust covariance and correlation coefficient estimator that resists against the presence of outliers in the dataset.

A robust estimator of covariance and correlation of bivariate random variables where introduced by Gnanadesikan and Kettenring (1972), that can be computed by substituting any robust scale estimator. On the observation that the classical correlation coefficient estimators have very low breakdown point, Abdullah (1990) proposed a robust correlation coefficient estimator on the basis of Least Median of Squares (LMS). The empirically proved breakdown value of this LMS based correlation estimator is nearly 50%. Shevlyakov (1997) provided a study of robust correlation coefficient on a bivariate normal distribution. He found that the median based correlation coefficient attained closes to the true

[1]Some part of this chapter is based on Sajana and Sajesh (2019)

value correlation in a highly contaminated situations as well. Falk (1997) proposed a median based robust alternative to the sample covariance between two random variables X and Y called comedian and the corresponding correlation coefficient is termed as correlation median. It turns out that MAD is a special case of comedian. Li et al. (2006) and Shevlyakov and Smirnov (2011) presented an extensive review and comparison of robust correlation methods through Monte Carlo experiments.

Similar type of estimator which robustly measure the degree of relation between two random variables has been developed by Sajana and Sajesh (2019). Here a location free scale estimator is used to define the dependence among two random variables. The scope for location free robust covariance and correlation estimators are discussed by Falk (1997). The characteristics of proposed robust covariance is compared to classical correlation estimator and some other well known alternative robust correlation coefficient estimators by utilizing theoretical and empirical results. A benchmark dataset is adopted to investigate the efficiency of the proposed estimator.

3.2 Robust S_n Covariance Estimator

Median is the most extensively known robust estimator for location of a random variable X. Usually, it gives $[n/2]^{th}$ order statistic from n independent observations $x_1, x_2, ..., x_n$ of X when n is odd. In the case where n is even, median is the average of $[n/2]^{th}$ and $([n/2]+1)^{th}$ order statistic. It is clear that median posses the optimal breakdown point 50%.

Several median based scale estimators are available in literature. A very popular median based robust scale estimator is Median Absolute Deviation from median (MAD) raised by Hampel (1974) and he established that it is an approximation of M estimator of scale. The asymptotic variance and influence function of MAD was derived by Huber (1981). A detailed study on limit the-

orems and strong consistency of MAD has been developed by Hall and Welsh (1985). MAD has breakdown point which is equal to that of median. Despite of high breakdown value, MAD has only 37% Gaussian efficiency in symmetric distributions. More efficient alternative for MAD with 50% breakdown point is discussed by Rousseeuw and Croux (1993). A pairwise distance estimator $Q_n(X) = 2.2219 \ \{|x_i - x_j| ; i < j\}_{(k)}$ where $k \approx \frac{\binom{n}{2}}{4}$ is one of the alternative to MAD. Another reliable substitute for MAD is

$$S_n(X) = 1.1926 \underset{i}{med} \underset{j}{med} |x_i - x_j|$$

where *med* stands for low median ($[\frac{n+1}{2}]^{th}$ order statistic) for outer median and high median ($([\frac{n}{2}] + 1)^{th}$ order statistic) for inner median and 1.1926 is the consistency factor for normal distributions. S_n estimator of scale assures bounded influence function and optimal breakdown point 50%. Even though S_n is less efficient than Q_n, S_n is more applicable because of its low gross error sensitivity. Thus, S_n is more robust than Q_n (Rousseeuw and Croux 1993).

Thus, classical covariance estimator is defined as

Consider the bivariate random variable (X, Y) and let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be n independent observations of (X, Y). Then $\bar{X} = \sum_{i=1}^{n} X_i$ and $\bar{Y} = \sum_{i=1}^{n} Y_i$ are the sample means of X and Y respectively. Empirical covariance between X and Y is defined as

$$\widehat{COV(X, Y)} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \tag{3.1}$$

It is clear that, $\widehat{COV(X, Y)}$ is highly influenced by the presence of outliers which decreases its breakdown point $1/n$. Asymptotically it will become zero.

The location free covariance estimator proposed by Sajana and Sajesh (2019),

denoted as $S_nCov(X,Y)$ is defined as follows

$$S_nCov(X,Y) = \underset{i}{med}\ \{\underset{j\neq i}{med}\ [(x_i - x_j)\,(y_i - y_j)]\} \tag{3.2}$$

where $1 \leqslant i, j \leqslant n$ and med stands for low median ($[\frac{n+1}{2}]^{th}$ order statistic). The square of consistency factor (1.1926) of $S_n(x)$ can be multiplied to $S_nCov(X,Y)$ in order to get consistency at normal distribution. The repeated use of median was introduced by Tukey (1977) and it is applied in estimation of linear regression by Siegel (1982). Clearly, the defined robust covariance estimator is designed on the basis of repeated median idea. Due to lemma by Siegel (1982), repeated median values are bounded. Further properties of proposed estimators are discussed below.

Assume that $\{z_i = (x_i, y_i); i = 1, ...n\}$ are independent observations from a Euclidean space $\mathcal{X}$ with common distribution $L = F \times G$ (F and G are distribution functions of X and Y respectively). Define a kernel function $u : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $u(z_i, z_j) = (x_i - x_j)(y_i - y_j)$. Let T_1 and T_2 be sample medians. For each z, let $U(z) = T_1(H_z(t))$ and $\theta = T_2(H)$ where $H_z(t)$ and H are distribution functions $u(z, Z)$ and $U(Z)$ respectively. Here, θ be the covariance that need to estimate. For estimating θ, first estimate $U(z_i)$ by $\hat{U}(z_i) = T_1(H_{z_i,n-1}(t))$, where $H_{z_i,n-1}(t)$ is the empirical distribution of $\{u(z_i, z_j); j \neq i, i\ fixed\}$. Then put $\hat{\theta}_n = T_2(H_n)$, where H_n is the empirical distribution of $\hat{U}(z_1), ..., \hat{U}(z_n)$.

Now, define the distribution function $H_z(t)$ i.e,

$$H_z(t) = P(u(z,Z) \leq t) = 1 - \int_{-\infty}^{x}\int_{-\infty}^{y-\frac{t}{x-x'}} l(x',y')d(y')d(x') - \int_{x}^{\infty}\int_{y-\frac{t}{x-x'}}^{\infty} l(x',y')d(y')d(x') \tag{3.3}$$

where l is the density function of L

Lemma 1. *If X and Y are independent and continuous with $F^{-1}(0.5) = G^{-1}(0.5) = 0$, then $S_nCov(X,Y) = 0$*

Proof. Since X and Y are independent

$$H_z(t) = 1 - \int_{-\infty}^{x} G(y - \frac{t}{x - x'})dF(x') - \int_{x}^{\infty} [1 - G(y - \frac{t}{x - x'})]dF(x') \quad (3.4)$$

Here, $U(z)$ solves for $H_z(U) = 0.5$. By (3.4), and since

$$sgn(F(t) - 0.5) = sgn(G(t) - 0.5) = sgn(t)$$

where $sgn(t) = -1$ if $t < 0$, $= 0$ if $t = 0$ and $= 1$ if $t > 0$

$$\begin{aligned} H_z(0) = 1 - F(x)G(y) - (1 - F(x))(1 - G(y)) &< 0.5, if\ sgn(xy) > 0 \\ &= 0.5, if\ sgn(xy) = 0 \\ &> 0.5, if\ sgn(xy) < 0 \end{aligned}$$

Hence, based on the similar arguments that of Hössjer et al. (1992), $sgn(U(z)) = sgn(xy)$.

Also

$$H(0) = P(sgn(xy) \leq 0) = 1 - F(0)G(0) - (1 - F(0))(1 - G(0)) = 0.5$$

This follows,

$$H^{-1}(0.5) = \underset{Z \sim L}{med}\ U(Z) = 0$$

$$S_n Cov(X, Y) = 0$$

□

Considering the two equalities in lemma 1.1 and 1.3 by Falk (1997), it is clear that $S_n(aX + b) = |a|S_n(X)$ and $S_n Cov(X, Y) = aS_n(X)^2$ when $Y = aX + b$ where $a, b \in \mathbb{R}$. Let $X = Y$, $S_n Cov(X, X) = S_n(X)^2$, this states that S_n is a special case of $S_n Cov$. Moreover $S_n Cov$ is symmetric, location invariant and

scale equivariant, i.e.

$$S_nCov(X, aY + b) = aS_nCov(X, Y) = aS_nCov(Y, X)$$

A robust location free alternative to the coefficient of correlation $\rho = \frac{COV(X,Y)}{\sigma_x \sigma_y}$ is therefore the S_n *correlation* is denoted by $\xi(X, Y)$ is defined as

$$\xi = \xi(X, Y) = \frac{S_nCov(X, Y)}{S_n(X)S_n(Y)}$$

By lemma 1, if X and Y are independent and symmetric $\xi(X, Y) = 0$. Similarly in the case where there is complete dependence i.e $Y = aX + b$ for bivariate normal random variable $\xi(X, Y) = sgn(a),\ almost\ surely$. Hence, $\xi \in \{-1, 1\}$.

3.3 Performance Analysis

S_n *correlation* is using simulation study. Also the method is applied on real life dataset to check the efficiency of the proposed method in real life situation.

3.3.1 Simulation Study

Croux and Dehon (2010) compared nonparametric correlation coefficient estimation using finite sample variances. Finite sample efficiencies are estimated through Mean Square Error (MSE) and it is defined as

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(\hat{\rho} - \rho)^2 \tag{3.5}$$

where $\hat{\rho}$ is the estimated correlation coefficient.

The same parameter is adopted to compare the proposed method with other similar method. S_n *correlation* is compared with Gnanadesikan-Kettenring correlation coefficient proposed by Gnanadesikan and Kettenring (1972), correlation

median suggested by Gnanadesikan and Kettenring (1972) and classical correlation coefficient. Gnanadesikan and Kettenring (1972) defined the robust correlation coefficient as

$$r_{\tilde{\sigma}} = \frac{\tilde{\sigma}^2(v_1) - \tilde{\sigma}^2(v_2)}{\tilde{\sigma}^2(v_1) + \tilde{\sigma}^2(v_2)}$$

where $v_1 = (X/\tilde{\sigma}(X) + Y/\tilde{\sigma}(Y))/\sqrt{2}$, $v_2 = (X/\tilde{\sigma}(X) - Y/\tilde{\sigma}(Y))/\sqrt{2}$ and $\tilde{\sigma}$ is the robust estimators of scale.

Shevlyakov and Smirnov (2011) substituted MAD, S_n and Q_n for $\tilde{\sigma}$ in Gnanadesikan-Kettenring correlation estimator and compare the efficiencies. The corresponding robust estimators of correlation coefficients are denoted by r_{MAD}, r_{S_n} and r_{Q_n}. This study includes these three estimates of robust correlation coefficient for the comparison.

The following simulation aims to give a performance evaluation of robust correlation coefficient defined using proposed covariance estimator. The MSE of proposed estimator is compared with the robust correlation coefficients discussed in Shevlyakov and Smirnov (2011), correlation median established by Falk (1997) and the classical coefficient of correlation r.

The simulation is preformed for $k = 10000$, $\rho = 0.8$ and different sample sizes n. The results are presented in Table 3.1, Table 3.2 and Table 3.3. Table 3.1 shows the empirical $n*$ MSEs of various estimators for datasets without outliers. The simulation is performed with varying amount of contamination and the results for 10% and 30% contamination is presented in Table 3.2 and Table 3.3 respectively. The results are similar in other cases as well. Table 3.1 shows that error of ξ is less for small sample sizes as compared to correlation median. On small and large sample sizes, classical coefficient of correlation r is the best for symmetric samples. From Table 3.2, it is clear that in contaminated situation the proposed method perform better than the other methods when $n > 50$ for large percentage of contamination.

Table 3.1: $n*$MSE in Symmetric distribution

	n				
	20	50	100	200	1000
Correlation median	1.825	2.17	2.465	4.040	10.333
ξ	1.200	1.54	2.241	5.431	16.312
r_{MAD}	0.778	0.502	0.432	0.389	0.366
r_{S_n}	0.462	0.293	0.246	0.238	0.248
r_{Q_n}	0.326	0.206	0.176	0.175	0.164
r	0.173	0.151	0.135	0.137	0.133

Table 3.2: $n*$MSE in Symmetric distribution with 10% outlier

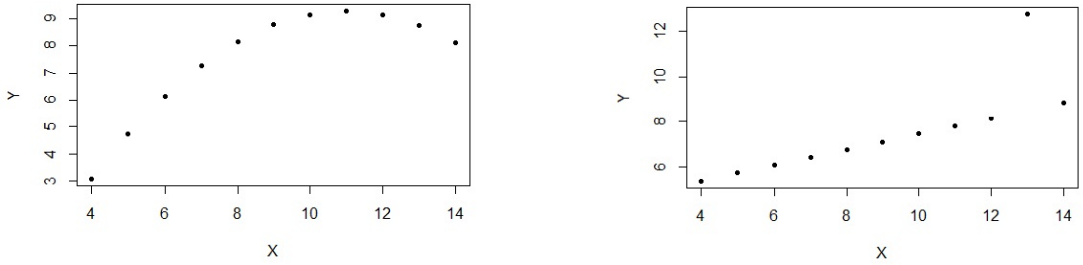
	n				
	20	50	100	300	1000
Correlation median	1.427	1.532	1.721	1.898	5.123
ξ	0.859	0.951	1.143	1.585	3.648
r_{MAD}	0.661	0.418	0.398	0.459	1.729
r_{S_n}	0.432	0.276	0.331	0.546	2.906
r_{Q_n}	0.278	0.236	0.375	0.732	3.875

Table 3.3: $n*$MSE in Symmetric distribution with 30% outlier

	n				
	20	50	100	300	1000
Correlation median	0.788	1.190	1.485	2.444	5.673
ξ	0.741	0.745	0.767	0.889	1.734
r_{MAD}	0.529	0.598	1.176	4.345	17.038
r_{S_n}	0.439	0.643	1.416	5.330	19.645
r_{Q_n}	0.288	0.608	1.393	4.857	17.203

3.3.2 Example

To understand the performance of correlation coefficient based on S_nCov, estimation is conducted for real life dataset. Anscombe's data consists of four bivariate dataset with different pattern of association has been described by Anscombe (1973) and the dataset is available *R package*. Two datasets consists of 12 observations are chosen from Anscombe data are used for evaluate the proposed estimator in real situation. These data consists of 12 observations, the scatter plot for these datasets are presented in Figure 3.1. The proposed correlation coefficient estimates are of datasets a and b are 0.35 and 1 respectively.



(a) (b)

Figure 3.1: Scatter plot of Anscombe Data

3.4 Summary

An efficient alternative for robust covariance estimator on the basis of repeated median is investigated. Covariance based on S_n estimator motivates to realize that it is a nested L-estimator. A non parametric measure of covariance and coefficient of correlation are proposed through this chapter. The $S_nCov(X, Y)$ satisfies characteristic of sample covariance in independent and symmetric random variables. The proposed estimator assures location invariant and scale equivariant properties as well. The efficiency of S_n *correlation* is greater than median correlation in terms of MSE. The error estimate of S_n *correlation* is lower than that of r_{Q_n} for large samples in contaminated situations. The proposed correla-

tion estimator is also applied in real dataset to prove the robustness of the estimator. This specifies that S_n *correlation* is more resistant than other estimates to the presence of outliers in the large sample cases. The direct generalization of proposed method to higher dimension is not be possible as the corresponding covariance matrix may not be positive definite. But an orthogonalization similar to that of developed by Maronna and Zamar (2002) may be helpful in *higher dimensional cases.

Chapter 4

Multidimensional Outlier Detection and Robust Estimation Using S_n Covariance[1]

4.1 Introduction

A multivariate outlier is an inconsistent combination of measurements of more than one variable. Application of univariate outlier detection methods in multivariate data may identify extreme observations in individual variables. An extensive use of univariate method to detect multivariate outliers may not be adequate, since it does not take in to account the relation among variables. In order to detect multivariate outliers, the distance from the center of mass and covariance structure must be equally considered. Mahalanobis Distance (MD) established by Mahalanobis (1936) is a multivariate measure of distance which consider deviation of observation from mean vector.

Mean vector and dispersion matrix are the only components of MD. Maximum likelihood estimates of these parameters are sensitive to the presence of outliers in the dataset. Hence substitution of these estimate in MD is inappropriate for outlier detection. For the purpose of outlier detection a Robust Mahalanobis Distance (RMD) is to be produced by employing robust estimates of the location and scatter parameters. Various methods have been introduced

[1]Some part of this chapter is based on Sajana and Sajesh (2020)

for robust estimation of location and dispersion of multivariate data in literature. Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) are introduced by Rousseeuw (1985) are widely accepted methods for robust estimation. MVE is based on the computation of minimum volume ellipsoid containing at least $h = [n/2] + 1$ of the observations of the data, where n is the number of samples. MCD searches for smallest covariance determinant which encompasses at least half of the data points. FAST-MCD was proposed by Rousseeuw and Van Driessen (1999) as an improved version of MCD. But it still needs substantial time for detection when the number of dimensions are more. Peña and Prieto (2001) and Peña and Prieto (2007) established Kurtosis algorithm for robust estimation. This method consists of maximization and minimization of projection of kurtosis coefficients based on some directions generated using stratified random sampling. This procedure also has some limitations in high dimensions and correlated samples. Orthogonalized Gnanadesikan-Kettenring (OGK) estimator introduced by Maronna and Zamar (2002) used a robust covariance matrix defined by Gnanadesikan and Kettenring (1972) which is non-positive semi definite and not an affine-equivariant. This method contains an orthogonalization technique which makes the covariance matrix positive definite and affine-equivariant. Similar type of orthogonalization is adopted by Sajesh and Srinivasan (2012) in the Comedian approach for multivariate outlier detection. In the context of psychological science, Leys et al. (2018) proposed a Robust Variant Mahalanobis Distance (RVMD) method for multivariate outlier detection. The RVMD method constitutes the MCD with two threshold values for detecting multivariate outliers in the data and both act differently for various contamination levels. An extension of median into a multidimensional situation is applied by Sajana and Sajesh (2018b) in detecting multivariate data. They proposed a multivariate outlier detection using Spatial Median(SM) and the performances of method is limited to small percentage of contamination in

the data. The recommended ratio of n and p is $n > 5p$ for the performance of MCD, in order to rectify this restriction, Boudt et al. (2019) proposed Minimum Regularized Covariance Determinant(MRCD) method that regularize $h-$ subset based on a predetermined positive definite target matrix.

In this chapter, a robust distance based approach is proposed using RMD. A multivariate version of the robust S_nCov proposed by Sajana and Sajesh (2019) and variable wise median are used for the computation of RMD (Sajana and Sajesh, 2020). The efficiency of proposed outlier detection method is measured through simulation studies. Robustness properties of this method is tested using theoretical and empirical approaches. Methods which are popularly known for multivariate outlier detection like Comedian, Kurtosis, FAST-MCD, OGK, SM method, RVMD and MRCD are compared with the proposed method.

4.2 Multidimensional Expansion of S_nCov

Let $\mathbf{X}$ be a $n \times p$ data matrix with independent observations $\mathbf{x}_i^T = \{x_1, ..., x_n\}$ and columns $\mathbf{X}_j (j = 1, ...p)$ the covariance matrix based on S_nCov is defined as

$$\mathbf{COV}_{S_n}(\mathbf{X}) = (S_nCov(X_i, X_j)) \ , \ i, j = 1, 2, ..., p \tag{4.1}$$

Corresponding correlation matrix of $\mathbf{COV}_{S_n}$ denoted by $\boldsymbol{\xi}_{S_n}$ is defined as

$$\boldsymbol{\xi}_{S_n}(\mathbf{X}) = \mathbf{DCOV}_{S_n}(\mathbf{X})\mathbf{D}^T \tag{4.2}$$

where $\mathbf{D}$ is diagonal matrix with diagonals $1/S_n(x_i), \ i = 1, ..., p$

Since S_nCov is a robust alternative for classical bivariate covariance, it is possible to state that $\mathbf{COV}_{S_n}$ is a robust alternative to covariance matrix. Basically, this matrix is non-positive semi definite. In order to solve non-positive semi definiteness, a procedure implemented by Maronna and Zamar (2002) to obtain positive definite and approximately affine equivariant scatter estimates is

adopted. To obtain positive definite dispersion matrix and robust estimates, the following steps are applied.

1. Define matrix $\mathbf{E}$ with columns $\mathbf{e}_j$ for $j = 1, ..., p$, where $\mathbf{e}_j$ is the eigenvector corresponding to eigenvalue λ_j of correlation matrix $\boldsymbol{\xi}_{S_n}$. Hence $\boldsymbol{\xi}_{S_n}$ can be written as $\boldsymbol{\xi}_{S_n} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^T$, where $\boldsymbol{\Lambda} = diag(\lambda_1, ..., \lambda_p)$

2. Let $\mathbf{R} = \mathbf{D}^{-1}\mathbf{E}$ and $\mathbf{z}_i = \mathbf{R}^{-1}\mathbf{x}_i$. Then assume that $\mathbf{z}_i^T (i = 1, ..., n)$ and $\mathbf{Z}_j (j = 1, ..., p)$ are rows and columns of orthogonalized matrix $\mathbf{Z}$.

3. The resulting robust estimates for location vector $\mathbf{L}_r(\mathbf{X})$ and scatter matrix $\mathbf{C}_r(\mathbf{X})$ in the following way,

$$\mathbf{m}_r(\mathbf{X}) = \mathbf{R}\boldsymbol{\nu} \tag{4.3}$$

$$\mathbf{C}_r(\mathbf{X}) = \mathbf{R}\boldsymbol{\Gamma}\mathbf{R}^T \tag{4.4}$$

where $\boldsymbol{\nu} = (med(\mathbf{Z}_1), ..., med(\mathbf{Z}_p))^T$ and $\boldsymbol{\Gamma} = diag(S_n(\mathbf{Z}_1)^2, ..., S_n(\mathbf{Z}_p)^2)$ here *med* stands for median and S_n is the robust scale estimate. This process can be iterated to improve estimates by replacing $\boldsymbol{\xi}_{S_n}$ with the form of $\mathbf{C}_r$.

Then squared RMD on the basis of robust estimates is defined as

$$RMD(\mathbf{x}_i, \mathbf{m}_r, \mathbf{C}_r) = rmd_i = (\mathbf{x}_i - \mathbf{m}_r)^T \mathbf{C}_r^{-1} (\mathbf{x}_i - \mathbf{m}_r),\ i = 1, ..., n \tag{4.5}$$

where $\mathbf{m}_r$ and $\mathbf{C}_r$ are defined in (4.3) and (4.4) respectively. Decision regarding cutoff value is one of the significant task in outlier detection. In order to increase the performance of proposed method an adjusted cutoff is considered for different regions of dimensions i.e,

$$cv = \begin{cases} b\chi^2_{(0.95,p)} & if\ p < 15 \\ \dfrac{\chi^2_{(0.95,p)} med(rmd_1, ..., rmd_n)}{\chi^2_{(0.5,p)}} & if\ p \geqslant 15 \end{cases} \tag{4.6}$$

Thus, an $\mathbf{x}_i$ observation is identified as an outlier if $RMD(\mathbf{x}_i, \mathbf{m}_r, \mathbf{C}_r) > cv$. A positive definite and robust estimate can be formulated by a weight function on the basis of RMD and cv. Here b is a constant which takes value 1 if $p \leq 5$ and 2.5 if $p > 5$. The resulting method of multivariate outlier detection using $\mathbf{COV}_{S_n}$ can be represented as S_n method of outlier detection.

4.3 Simulation

The effectiveness of proposed S_n method is tested through a series of simulation processes and later experimented with real datasets. Masking and swamping are the two errors occurring while detecting possible outliers. Two aspects of outlier identification are assayed i.e, rate of successful complete detection of contained outliers which is expressed by Rate of Successful Detection (RSD) and Rate of False Detection (RFD) indicating rate of detection of inliers as outliers. Sajesh and Srinivasan (2012) presented Comedian method and found out that it is better than Kurtosis, FAST-MCD and OGK with RSD and RFD. In this article, the proposed method is compared with Comedian and other methods.

To create a data contaminated with outliers, $100(1-\gamma)$ observations are generated from $N(\mathbf{0}, \mathbf{I})$ distribution with dimension p for a given level of contamination γ and 100γ observations are replaced by $N(\delta\mathbf{a}, \lambda\mathbf{I})$ distribution, where $\mathbf{a}$ represents the vector $(1, ..., 1)^T$ and $\mathbf{I}$ the identity matrix. The test is undertaken for different choices of dimensions p ($p = 5, 10, 20$) and contamination level γ ($\gamma = 0.1, 0.2, 0.3$). For determining the ability to identify minor disparities in data, the experiment is performed for small deviations of δ ($\delta = 5, 10$) and λ ($\lambda = 0.01, 0.25, 1$).

Table 4.1 and Table 4.2 shows the RSD values of S_n, Comedian, Kurtosis, FAST-MCD, OGK, SM method, RVMD and MRCD for $\delta = 5$ and $\delta = 10$ respectively. Some comparable situation of Comedian method presented by Sajesh and Srinivasan (2012) is chosen to produce this table. The rates from the table

shows that S_n method works better than Comedian and Kurtosis apart from two cases ($p = 10, \gamma = 0.3, \delta = 5, \lambda = 1$ and $p = 20, \gamma = 0.3, \delta = 5, \lambda = 0.01$). Table 4.3 and Table 4.4 exhibits maximum RFD values in all combinations, comparison of S_n method with other outlier detection methods for location sifts $\delta = 5$ and $\delta = 10$ respectively. All the combinations of values explained in simulation part are considered for the maximum RFD estimation. In all the cases, S_n method performed better than rest of the methods with zero RFD.

Table 4.1: RSD Comparison

			$\delta = 5$							
p	λ	γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
5	0.25	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	100	100	100	100	99	100
		0.3	95	95	98	60	81	86	0	0
	0.01	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	99	39	100	100	99	100
		0.3	99	70	99	0	34	94	0	0
	1	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	98	100	100	100	100	100
		0.3	100	100	97	100	83	71	0	0
10	0.25	0.1	100	100	100	100	00	100	100	100
		0.2	100	100	100	41	100	100	57	100
		0.3	100	99	79	0	99	41	0	0
	0.01	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	99	0	100	100	100	100
		0.3	100	83	91	0	38	56	0	0
	1	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	75	100	100	100	100	100
		0.3	97	99	21	99	100	51	0	0
20	0.25	0.1	100	100	100	95	100	100	100	100
		0.2	100	100	90	0	100	75	3	100
		0.3	100	100	3	0	100	0	0	0
	0.01	0.1	100	100	100	0	100	78	42	100
		0.2	100	100	85	0	100	81	0	100
		0.3	88	99	0	0	52	0	0	0
	1	0.1	100	100	49	100	100	52	100	100
		0.2	100	100	1	100	100	58	100	100
		0.3	100	100	0	2	100	0	97	0

4.3.1 Simulation in correlated data

The behavior of S_n method in correlated data is analyzed because of its lack of affine-equivariance. Devlin et al. (1981) applied a correlation matrix $\mathbf{P}$ of

Table 4.2: RSD Comparison

			$\delta = 10$							
p	λ	γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
5	0.25	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	100	100	100	100	100	100
		0.3	100	100	100	100	100	100	0	0
	0.01	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	100	100	100	100	100	100
		0.3	100	100	100	0	100	100	0	0
	1	0.1	100	100	46	100	100	100	100	100
		0.2	100	100	1	100	100	100	100	100
		0.3	100	100	0	100	100	100	0	0
10	0.25	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	100	100	100	100	100	100
		0.3	100	100	90	0	100	99	0	0
	0.01	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	100	0	100	100	74	100
		0.3	100	100	92	0	100	98	0	0
	1	0.1	100	100	100	100	100	100	100	100
		0.2	100	100	100	100	100	100	100	100
		0.3	100	100	38	100	100	99	0	0
20	0.25	0.1	100	100	100	100	100	52	100	100
		0.2	100	100	92	0	100	99	1	100
		0.3	100	100	2	0	100	30	0	0
	0.01	0.1	100	100	100	86	100	100	100	100
		0.2	100	100	94	0	100	100	0	100
		0.3	100	100	3	0	100	46	0	0
	1	0.1	100	100	46	100	100	100	100	100
		0.2	100	100	1	100	100	99	100	100
		0.3	100	100	0	2	100	24	97	0

dimension p ($p = 6$) for generating Monte Carlo data from different distributions. The correlation matrix $\mathbf{P} = ((\rho_{ij}))$ has the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 \end{bmatrix} \textit{ where } \mathbf{P}_1 = \begin{bmatrix} 1 & 0.95 & 0.30 \\ 0.95 & 1 & 0.10 \\ 0.30 & 0.10 & 1 \end{bmatrix}, \ \mathbf{P}_2 = \begin{bmatrix} 1 & -0.499 & -0.499 \\ -0.499 & 1 & -0.499 \\ -0.499 & -0.499 & 1 \end{bmatrix}$$

The dimension of correlation matrix is large enough to study the multivariate estimate. Here the range of correlation is high that helps to investigate the capability of this method to identify the outliers in highly correlated dataset. Asymmetrical datasets of size 100 is generated which includes $100(1-\gamma)$ observation from $N(\mathbf{0}, \mathbf{P})$ and 100γ observation from $N(5\mathbf{a}, \mathbf{P})$, where $\mathbf{a} = (1, ..., 1)^T$. The RFD values for proposed method in correlated data are presented in Table 4.5. The results in Table 4.5 shows that, RFD of S_n method is zero i.e the

Table 4.3: RFD Comparison

			$\delta = 5$							
λ	p	γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
0.01	5	0.1	0	4	7	17	15	0	15	15
		0.2	0	5	7	41	10	0	23	5
		0.3	0	4	5	51	10	0	31	25
	10	0.1	0	4	6	22	16	0	60	15
		0.2	0	6	42	44	13	0	68	5
		0.3	0	3	45	54	7	0	70	25
	20	0.1	0	1	10	39	18	0	90	15
		0.2	0	3	40	47	12	0	80	5
		0.3	0	3	40	63	12	5	70	15
0.25	5	0.1	0	3	5	18	16	0	15	15
		0.2	0	2	5	11	11	0	13	5
		0.3	0	2	5	32	7	0	25	25
	10	0.1	0	2	5	24	20	0	61	15
		0.2	0	2	7	36	10	0	67	5
		0.3	0	2	31	40	9	0	62	25
	20	0.1	0	1	9	38	16	0	90	15
		0.2	0	2	13	39	10	0	80	5
		0.3	0	1	39	40	7	0	70	19
1	5	0.1	0	3	6	14	15	0	15	15
		0.2	0	2	6	9	13	0	13	5
		0.3	0	2	6	7	7	0	13	25
	10	0.1	0	2	9	23	16	0	59	15
		0.2	0	2	6	15	13	0	50	5
		0.3	0	2	6	13	7	0	44	25
	20	0.1	0	2	8	28	16	0	90	15
		0.2	0	1	5	18	12	0	80	5
		0.3	0	1	4	27	8	0	70	7

proposed method is free from false detection of inliers as outliers.

4.3.2 Equivariance

This section discuss about the equivariance property of proposed method by simulated data. Equivariance study is significant to the proposed method as the initial estimate of dispersion is not equivariant. Consider a multidimensional random variable $\mathbf{X} = \{\mathbf{x}_1, ...\mathbf{x}_n\}$ with each $\mathbf{x} \in \mathbb{R}^p$. Let $\mathbf{X_A} = \{\mathbf{Ax}_1, ...\mathbf{Ax}_n\}$ where $\mathbf{A}$ is $p \times p$ nonsingular matrix. If the estimates of location and scatter are affine-equivariant, then

$$\boldsymbol{m_A = m(X_A) = Am(X)} \text{ and } \boldsymbol{C_A = C(X_A) = AC(X)A^T}$$

Table 4.4: RFD Comparison

			$\delta = 10$							
λ	p	γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
0.01	5	0.1	0	3	9	18	11	0	16	15
		0.2	0	6	7	12	12	0	14	5
		0.3	0	4	5	47	7	0	33	25
	10	0.1	0	2	7	23	19	0	59	15
		0.2	0	5	5	42	11	0	67	5
		0.3	0	5	45	56	9	0	68	2
	20	0.1	0	1	8	38	16	0	90	15
		0.2	0	3	40	47	14	0	80	5
		0.3	0	3	40	61	7	2	70	25
0.25	5	0.1	0	5	7	13	14	0	15	15
		0.2	0	3	6	12	10	0	13	5
		0.3	0	2	5	8	6	0	11	25
	10	0.1	0	2	8	21	17	0	46	15
		0.2	0	2	7	15	13	0	52	5
		0.3	0	4	24	37	7	0	61	6
	20	0.1	0	1	8	28	19	0	90	15
		0.2	0	1	14	39	12	0	80	5
		0.3	0	1	40	41	9	0	70	25
1	5	0.1	0	3	6	19	14	0	15	15
		0.2	0	2	6	10	11	0	13	5
		0.3	0	2	6	7	8	0	11	2
	10	0.1	0	2	6	23	18	0	63	15
		0.2	0	3	6	18	10	0	44	5
		0.3	0	1	7	9	7	0	47	25
	20	0.1	0	2	9	28	14	0	90	15
		0.2	0	1	5	19	11	0	80	5
		0.3	0	1	4	30	7	0	70	25

Table 4.5: RFDs of S_n method in correlated samples

γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
0.1	0	7	10	19	17	2	18	25
0.2	0	4	4	14	11	0	19	25
0.3	0	2	5	7	6	1	25	25

The Mahalanobis distance of $\mathbf{X_A}$ from $\mathbf{m_A}$ based on $\mathbf{C_A}$ holds affine-equivariance property if both the location and scatter are affine-equivariant. Maronna and Zamar (2002) generated a random matrices as $\mathbf{A} = \mathbf{TD}$ where $\mathbf{T}$ is a random orthogonal matrix and $\mathbf{D} = diag(u_1, ..., u_p)$, where u_j's are independent and uniformly distributed in $(0, 1)$.

Simulation of untransformed data has been repeated to investigate the performance of proposed method under transformation. Each data matrix is transformed by multiplying random non-singular matrix. The proposed method is then applied to the transformed data matrix to detect outlier. The experiment is conducted to different values of p ($p = 5, 10, 20$) and contamination level γ ($\gamma = 0.1, 0.2, 0.3$). Table 4.6 shows simulated results under transformed data and it could be observed that the S_n method is able to detect all the outliers in the dataset, except for some stray situations.

4.3.3 Breakdown value of S_n method

Maximum proportion of outlier that an estimator can safely tolerate before giving incorrect estimate is termed as breakdown value. Similarly, the breakdown value of an outlier detection method could be defined as the maximum proportion (γ^*) of outliers that the method can precisely identify. Clearly, if $\gamma > \gamma^*$ the method fails to detect majority of the outliers and faultily spot the inliers as outliers or decreases RSDs and increases RFDs. Hence, it is relevant to use RSD and RFD for examining the breakdown value of an outlier detection method.

The experiment to find the breakdown value of S_n method contains generation of symmetrically and asymmetrically distributed contaminations. At first, data of size n is simulated from $N(\mathbf{0}, \mathbf{I})$ with dimension p. Then symmetric outliers are inserted by multiplying i_{th} observation with $100i$. For asymmetric contamination i_{th} observation was replaced by $(100i)\mathbf{1}$, where $\mathbf{1} = (1, ...1)$. Different values of p ($p = 10, 30, 50, 80, 100$) and γ ($\gamma = 10, 20, 30, 40, 48$) were chosen to determine empirical breakdown value of S_n method. The results for selected sample size $n = 1000$ are presented in Table 4.7. This empirical experiment shows 100% RSD and 0 RFD.

Table 4.6: RSDs and RFDs of S_n method in transformed data

			$\delta = 5$		$\delta = 10$	
λ	p	γ	RSD	RFD	RSD	RFD
0.01	5	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	99	0	100	0
	10	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
	20	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
0.25	5	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	88	0	100	0
	10	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
	20	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	84	0	100	0
1	5	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
	10	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
	20	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0

4.4 Real Dataset

The efficacy of proposed method in real dataset is explained in this section. Bushfire data is considered for studying real data application and it was collected by Campbell (1989) which consist of satellite measurement on 5 frequency bands each corresponding to 38 pixels. The Bushfire dataset is also openly available at https:// vincentarelbundock.github.io/ Rdatasets/datasets.html. Maronna and

Table 4.7: Empirical results for breakdown value

		Symmetric		Asymmetric	
p	γ	RSD	RFD	RSD	RFD
10	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0
30	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0
50	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0
80	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0
100	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0

Yohai (1995) analyzed the dataset and concluded that observations 7-11 are outlying and they can be easily identified by various robust methods. But, the observations 32-38 are masked by the first group of outliers and Stahel-Donoho projection estimator implemented by Maronna and Yohai (1995) does not get effected by this error. Outliers in bushfire data are identified using S_n method, Comedian, Kurtosis, FAST-MCD, OGK, SM method, RVMD and MRCD and the diagnostic plot is presented in Figure 4.1. From the figure, it can be see that the S_n method is able to detect observations 7-11 and 31-38 as possible outliers

and it has relatively better result with less swamping error. In the case of other methods, Comedian method identified 8-9 and 30-38 as outliers. kurtosis method observed that sample 30 and FAST-MCD method indicated that observation 29 are additional outliers. According to OGK method, it is found that sample 28 is also a deviated observation. In addition to S_n method, SM method is able to detect the possible outliers. But RVMD and MRCD are only capable of detecting few outliers presented in the dataset.

4.5 Summary

Outlier detection is a significant part of data preprocessing since it could influence the inferences of analysis. An alternative method to detect multivariate outliers on the basis of repeated median covariance matrix is presented through this chapter. The effectiveness of the method is discussed and compared with well-known methods comedian, kurtosis FAST-MCD, OGK, SM method, RVMD and MRCD.

The simulation study is executed and explained in different possible choices of parameters. Simulation results of RSD and RFD shows that the proposed method performed better than Kurtosis, FAST-MCD, SM method, RVMD and MRCD. In the case of comparison with comedian and OGK, the proposed method appeared better in RSD measurements except some rare cases. But it outperformed in RFD values. To understand the capability of proposed method in collinear data, highly correlated data is generated in specific dimension. The RFDs presented here reflects low swamping error of S_n method in correlated data. Affine-equvariance property of the method is also tested beacause of lack of equivariance of $\mathbf{COV}_{S_n}$. RSDS and RFDS seems similar in both affinely transformed and untransformed data. Symmetrically and asymmetrically contaminated datasets are generated to estimate the breakdown value of proposed method. The simulation result of breakdown value shows that, the method is ro-

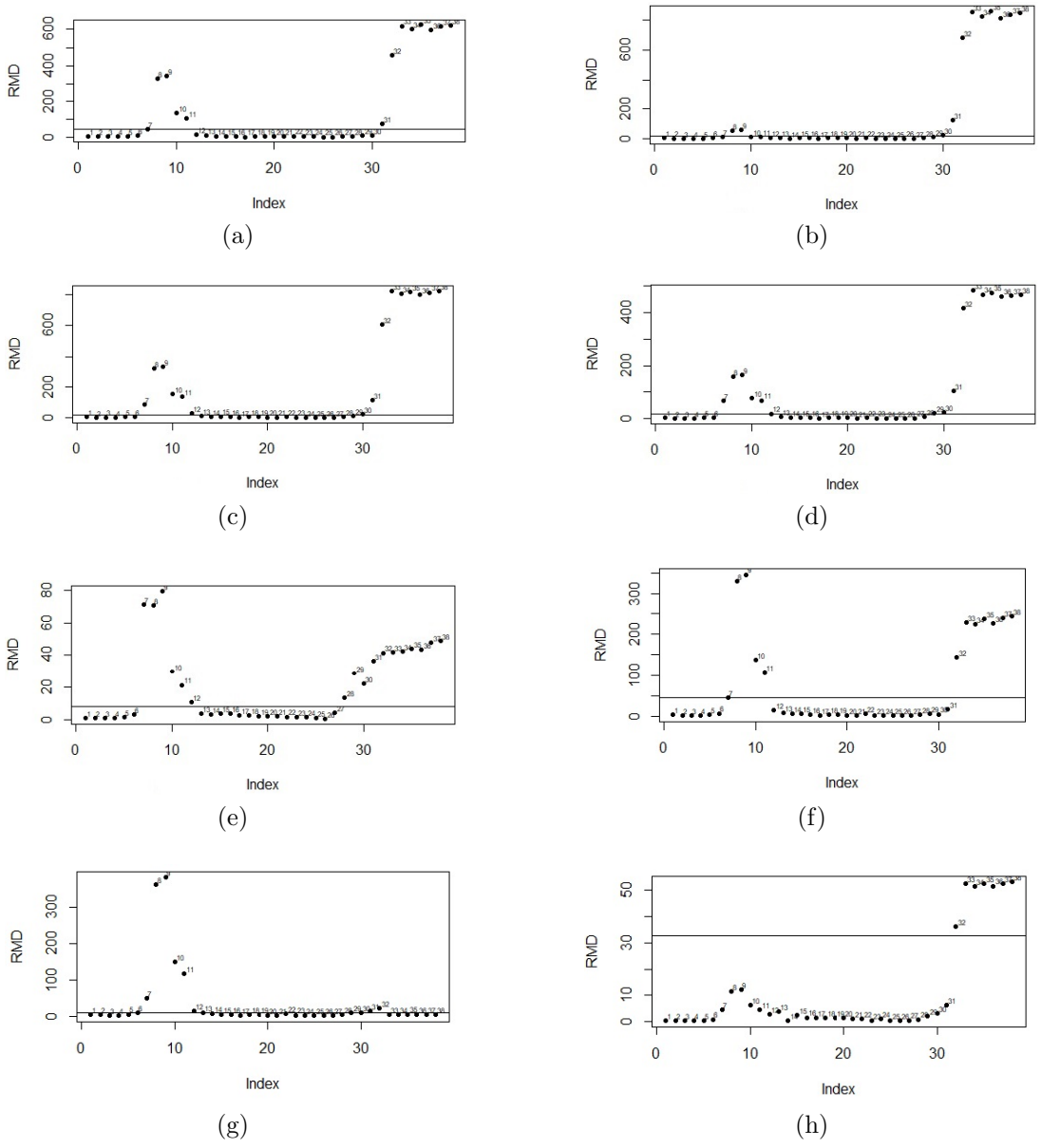


Figure 4.1: Outlier detection plot for Bushfire data.(a) S_n method, (b) Comedian, (c)Kurtosis, (d) FAST-MCD, (e) OGK, (f) SM, (g) RVMD, (h) MRCD

bust even under highly contaminated situations. In the real datasets SM method performed equivalent to S_n method, but performance of SM method in simulated datasets are less uncompromisable. The *R Programming code* for performing S_n method is provided in the Appendix B. The application of proposed method in real dataset reflects its effectiveness of simulated result by detecting possible outliers with low swamping. Hence, S_n method can apply in multivariate datasets for cleansing multiple outliers with minimum errors.

Chapter 5

Robust Quadratic Discriminant Analysis Using S_n covariance

5.1 Introduction

Discriminant Analysis (DA) is the multivariate technique that allows separating random objects into known groups of the population. The theory of discriminant function was introduced by Fisher (1936) for implementing the treatment of multiple measurements. The discriminant analysis can be considered as a statistical decision-making problem (Anderson 2004). The objective of discriminant analysis is the formulation of classification rules based on several training dataset and these determined rules are applied to classify the actual dataset. Discriminant analysis method includes Linear Discriminant analysis (LDA) and Quadratic Discriminant Analysis (QDA) for the assumptions according to equal and unequal population covariance matrices.

The classical methods of discriminant rules are often adopted to allocate multivariate observations to population groups and these are functions of sample mean vector and covariance matrix of the training dataset. Unfortunately, traditional rules are influential to outlying observations in the dataset which can mislead the classification of actual data. To overcome this situation, a robust alternative that is less sensitive to the presence of outlying observations are required for the estimation of parameters of discriminant rules.

Several multivariate robust estimation methods have been applied in literature for constructing robust quadratic discriminant rules. Robust quadratic and linear discriminant analysis using MCD estimators was investigated by Hubert and Van Driessen (2004). They showed that the reweighting technique applied in MCD decreases the misclassification probabilities.Kurtosis method proposed by Peña and Prieto (2001) was implemented by Lakshmi and Sajesh (2018) in robust estimation of QDA parameters. Sajesh and Srinivasan (2019) developed robust QDA using comedian method and presented that the method is better than that of robust QDA using MCD and classical QDA.

This chapter focuses on the study of Robust Quadratic Discriminant Analysis (RQDA) using the robust location and scatter based on S_n method discussed in the previous chapter. The effect of robust quadratic discriminant rules is investigated by comparing the overall misclassification estimate (MP) proposed by Hubert and Van Driessen (2004). The proposed robust QDA is compared with classical estimators and the RQDA's proposed by Hubert and Van Driessen (2004) and Sajesh and Srinivasan (2019), to test the efficiency of the method. Moreover, real data applications are illustrated to ensure the performance of proposed RQDA in real life situations.

5.2 Classical Quadratic Discriminant Analysis

The theoretical generalization of classification procedure for discrimination with several groups of population $\pi_1, ..., \pi_k$ can be explained by considering the density $f_i(\mathbf{x})$ associated with population π_i to be multivariate normal with mean vectors $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ (Johnson and Wichern 1992). The derived discriminant rule for allocating the multivariate observation $\mathbf{x} \in \mathbf{R}^p$ to l^{th} population group is defined as

allocate $\mathbf{x}$ to π_l if

$$\ln p_l f_l(\mathbf{x}) = \ln p_l - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_l| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l) \qquad (5.1)$$

where p_i be the membership probability of population group π_i. The above discriminant rule can be simplified by ignoring the constant term, then the quadratic discriminant score will be

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ \ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p_i \quad for\ i = 1, 2, ..., k \quad (5.2)$$

It is applied to find the discriminant rule with least total misclassification probability for normal population (Johnson and Wichern 1992). The discriminant rule is derived as
allocate $\mathbf{x}$ to π_l if

$$d_l^Q\mathbf{x} = max\{d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), ..., d_k^Q(\mathbf{x})\} \qquad (5.3)$$

The quadratic discriminant score is reduced for the homogeneous population covariance matrices, it will be a linear combination of components of $\mathbf{x}$. Therefore the linear discriminant score is defined as

$$d_i^L(\mathbf{x}) = \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \quad for\ i = 1, 2, ..., k \qquad (5.4)$$

Practically, the scores are embodied of unknown parameters, $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and p_i (membership probabilities). Thus, sample mean vector $\bar{\mathbf{x}}_i$ and sample covariance matrix $\mathbf{S}_i$ of training datasets are adopted to compute the discriminant score explained in equations 5.2. The estimated Classical Quadratic Discriminant Rule (CQDR) or QDA_C is then written as

allocate $\mathbf{x}$ to π_l if

$$d_l^{CQ}\mathbf{x} = max\{d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), ..., d_k^Q(\mathbf{x})\} \tag{5.5}$$

where $d_i^Q\mathbf{x}$ is defined as

$$\hat{d}_i^{CQ}(\mathbf{x}) = -\frac{1}{2}\ \ln|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \ln\hat{p}_i^C \quad for\ i = 1, 2, ..., k \tag{5.6}$$

The unknown membership probability can be estimated as a constant, *i.e* $\hat{p}_i^C = 1/k$ or can be estimated using relative frequencies of each population group, *i.e* $\hat{p}_i^C = n_i/n$ where $n = \sum_i^k n_i$. Since the classical discriminant function directly depends on the classical estimators of mean vector and covariance matrix of training data which are highly influenced by the presence of outliers, classification based on the classical discriminant function will be misleading. In order to solve the disparity in discrimination of observation due the presence of outliers, it is preferable to adopt robust estimators of mean vector and covariance matrix in the classification rule. The robust quadratic discriminant rule based on S_n estimators is described in the following section.

5.3 Robust Quadratic Discriminant Analysis (RQDA)

The robust quadratic discriminant rule for RQDA is then defined as,

allocate $\mathbf{x}$ to π_l if $\hat{d}_l^{RQ}(x) > \hat{d}_i^{RQ}(x)$ for all $i = 1, 2, ..., k$

$$\hat{d}_i^{RQ}(\mathbf{x}) = -\frac{1}{2}\ \ln|\hat{\mathbf{\Sigma}}_{i,S_n}| - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_{i,S_n})^T\hat{\mathbf{\Sigma}}_{i,S_n}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_{i,S_n}) + \ln\hat{p}_i^R \quad for\ i = 1, 2, ..., k \tag{5.7}$$

where $\hat{\boldsymbol{\mu}}_{i,S_n}$ and $\hat{\mathbf{\Sigma}}_{i,S_n}$ are the estimates of mean vector and covariance matrix using S_n method. The membership probability can be defined robustly by $\hat{p}_i^R = \tilde{n}_i/\tilde{n}$, where $\tilde{n} = \sum_{i=1}^n \tilde{n}_i$ and $\tilde{n}_i$ is the number of inliers in the i_{th} group. The performances of the RQDA based on S_n *method* (RQDA_{S_n}) is then evaluated

using estimated MP proposed by Hubert and Van Driessen (2004). The MP is defined as the weighted mean of the misclassification probabilities where weights are estimated membership probabilities.

$$MP = \sum_{i=1}^{k} \hat{p}_i^R \; MP_i \tag{5.8}$$

where MP_i be the misclassification probabilities. In this chapter the evaluation of robust discriminant rules are conducted using R-programming language. To ensure the performance of the proposed RQDA_{S_n} is compared with the classical discriminant analysis, RDA based on MCD estimator (Hubert and Van Driessen 2004) and the classical discriminant analysis, using simulated samples.

5.4 Simulation Results

The technique of MP includes splitting the observations randomly into two sets, one is the *training set* which is utilized for constructing discriminant rule and other set is the *validation set* which is used to estimate misclassification error. The estimated MP values for different case of contamination is discussed below

The case A_p considers the uncontaminated data with dimension p where 500 observations from each population are drawn as training, which is denoted by

$A_p.\pi_1 : 500 \; N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p})$

$\pi_2 : 500 \; N_p(\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{2,p})$

$\pi_3 : 500 \; N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p})$

Training datasets which also contain outliers are samples from another distribution. These cases are given below.

$B_p.\pi_1 : 400 \; N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) + 100 \; N_p(6\boldsymbol{\mu}_{1p}, \boldsymbol{\Sigma}_{4,p})$

$\pi_2 : 400 \; N_p(\boldsymbol{\mu}_{2p}, \boldsymbol{\Sigma}_{2p}) + 100 \; N_p(6\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{4,p})$

$\pi_3 : 400 \; N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) + 100 \; N_p(6\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{4,p})$

$C_p.\pi_1 : 800 \; N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) + 200 \; N_p(6\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{4,p})$

$\pi_2 : 600\ N_p(\boldsymbol{\mu}_{2p}, \boldsymbol{\Sigma}_{2,p}) + 150\ N_p(6\boldsymbol{\mu}_{1p}, \boldsymbol{\Sigma}_{4,p})$

$\pi_3 : 400\ N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) + \pi_3 : 100\ N_p(6\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{4,p})$

$D_p.\pi_1 : 800\ N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{\mathbf{1,p}}) + 200\ N_p(6\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{\mathbf{4,p}})$

$\pi_2 : 400\ N_p(\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{\mathbf{2,p}}) + 100\ N_p(6\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{4,p})$

$\pi_3 : 400\ N_p(\boldsymbol{\mu}_{3p}, \boldsymbol{\Sigma}_{3,p}) + 100\ N_p(6\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{4,p})$

$E_p.\pi_1 : 400\ N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) + 100\ N_p(6\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{4,p})$

$\pi_2 : 450\ N_p(\boldsymbol{\mu}_{2p}, \boldsymbol{\Sigma}_{2,p}) + 50\ N_p(6\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{4,p})$

$\pi_3 : 350\ N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) + 150\ N_p(6\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{4,p})$

$F_p.\pi_1 : 160\ N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) + 40\ N_p(6\boldsymbol{\mu}_{3,p}, \mathbf{25}\boldsymbol{\Sigma}_{4,p})$

$\pi_2 : 160\ N_p(\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{2,p}) + 40\ N_p(6\boldsymbol{\mu}_{1,p}, \mathbf{25}\boldsymbol{\Sigma}_{4,p})$

$\pi_3 : 160\ N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) + 40\ N_p(6\boldsymbol{\mu}_{2,p}, \mathbf{25}\boldsymbol{\Sigma}_{4,p})$

where $\boldsymbol{\mu}_{i,p}$ is the zero vector with i^{th} element equal to 1. The different choices of covariance matrix $\boldsymbol{\Sigma}$.

$\boldsymbol{\Sigma}_{1,3} = diag(0.4, 0.4, 0.4)^2$

$\boldsymbol{\Sigma}_{1,5} = diag(0.4, 0.4, 0.4, 0.4, 0.4)^2$

$\boldsymbol{\Sigma}_{1,10} = diag(0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4)^2$

$\boldsymbol{\Sigma}_{1,10} = diag(0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4)^2$

$\boldsymbol{\Sigma}_{2,3} = diag(0.25, 0.75, 0.25)^2$

$\boldsymbol{\Sigma}_{2,5} = diag(0.25, 0.75, 0.25, 0.75, 0.25)^2$

$\boldsymbol{\Sigma}_{2,10} = diag(0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75)^2$

$\boldsymbol{\Sigma}_{2,20} = diag(0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.2$

$\boldsymbol{\Sigma}_{3,3} = diag(0.9, 0.6, 0.3)^2$

$\boldsymbol{\Sigma}_{3,5} = diag(0.9, 0.6, 0.3, 0.9, 0.6)^2$

$\boldsymbol{\Sigma}_{3,10} = diag(0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9)^2$

$$\mathbf{\Sigma}_{3,20} = diag(0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9, 0.6)^2$$

where *diag* stands for diagonal elements. Different situations of data contaminations are constructed using 20% outliers in B_p, C_p, D_p, E_p and F_p. From these various cases of training datasets, the case B_p contains equal number of observations and outliers. In the case of populations C and D, an unequal group size is considered. A varying outlier percentages are tested in trial dataset E and F. Each case is repeated for 100 Monte Carlo simulations and based on the inliers identified by the S_n method is then used to calculate relative frequencies of membership probabilities.

Table 5.1 and Table 5.2 respectively shows average of total misclassification probabilities of 100 Monte Carlo experiments of each training groups and over all misclassification for $p = 10$ and $p = 20$. The results tabulated on both tables shows that the group wise misclassification probability and the over all misclassification of RQDA$_{S_n}$ is less in all cases compared to RQDA $_{MCD}$ and RQDA$_C$. In comparison with RQDA$_{Comedian}$, RQDA$_{S_n}$ have less misclassification measurements in most of the cases. In Table 5.2 the misclassification decreasing rate increases for increse in the dimension as compared to RQDA$_{Comedian}$.

5.5 Real Life Example

First example of Hemophilia data is considered to evaluate the performance of RQDA$_{S_n}$ in real life data. This data consists of measurements of two variables on 75 women which contains, 45 hemophilia A carriers and 30 normal women, where the first variable measures log(AHF activity) and second variable measures log(AHF-like antigen). Johnson and Wichern (1992) studied and analyzed the dataset.

To determine discriminant rules, 60% of randomly selected data points are considered as a training dataset an the robust covariance matrix is calculated.

Table 5.1: Misclassification probability of RQDA_{S_n}, $\text{RQDA}_{Comedian}$ RQDA_{MCD} and RQDA_C for $p = 10$

	RQDA_{S_n}				$\text{RQDA}_{Comedian}$			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.046	0.033	0.98	0.354	0.047	0.034	0.99	0.355
B_p	0.038	0.097	0.155	0.102	0.039	0.110	0.177	0.114
C_p	0.021	0.073	0.152	0.072	0.025	0.095	0.201	0.093
D_p	0.014	0.641	0.150	0.099	0.014	0.744	0.177	0.115
E_p	0.037	0.137	0.266	0.155	0.036	0.121	0.245	0.139
F_p	0.050	0.036	0.040	0.042	0.051	0.041	0.044	0.046
	RQDA_{MCD}				QDA_C			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.046	0.034	0.9845	0.355	0.044	0.031	0.039	0.38
B_p	0.207	0.097	0.082	0.129	0.142	0.074	0.088	0.101
C_p	0.157	0.089	0.111	0.124	0.134	0.344	0.377	0.285
D_p	0.063	0.452	0.092	0.092	0.126	0.797	0.141	0.355
E_p	0.201	0.160	0.300	0.207	0.151	0.056	0.213	0.141
F_p	0.038	0.046	0.058	0.047	0.001	0.324	0.824	0.384

Table 5.2: Misclassification probability of RQDA_{S_n}, $\text{RQDA}_{Comedian}$ RQDA_{MCD} and RQDA_C for $p = 20$

	RQDA_{S_n}				$\text{RQDA}_{Comedian}$			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.021	0.012	0.051	0.016	0.024	0.116	0.99	0.344
B_p	0.082	0.0303	0.109	0.078	0.88	0.020	0.091	0.070
C_p	0.044	0.052	0.147	0.075	0.051	0.041	0.131	0.069
D_p	0.009	0.442	0.089	0.064	0.016	0.381	0.097	0.066
E_p	0.011	0.011	0.043	0.025	0.071	0.025	0.152	0.084
F_p	0.039	0.011	0.018	0.023	0.042	0.018	0.018	0.026
	RQDA_{MCD}				QDA_C			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.024	0.013	0.996	0.344	0.022	0.009	0.014	0.015
B_p	0.276	0. 052	0. 027	0.121	0.094	0.016	0.025	0.045
C_p	0.196	0.092	0.039	0.131	0.083	0.321	0.348	0.251
D_p	0.084	0.782	0.032	0.110	0.064	0.842	0.085	0.330
E_p	0.302	0.187	0.313	0.250	0.101	0.012	0.152	0.088
F_p	0.036	0.031	0.032	0.033	0.012	0.481	0.754	0.411

The estimated membership probabilities using inliers in the selected training set are $p_1^R = 0.4$ and $p_2^R = 0.6$. The remaining 30% observations are considered as validation set to compute misclassification probability and MP. The estimated group misclassification using RQDA_{S_n} are 12%for group I, 8% for group II and over all misclassification 22%. The misclassification estimate is equivalent to that of CQDA since the data is uncontaminated.

5.6 Summary

Discriminant analysis is related to the discriminant score and classification rule associated with it. Since the classical discriminant scores are highly sensitive to the presence of outliers in the dataset, a more efficient RQDA is proposed in this chapter. The RQDA is constructed based on the robust estimation procedure developed on the basis of S_n method.

The evaluation of the performance of RQDA_{S_n} is conducted using Monte Carlo simulation study and it is compared with RQDA_{com}, RQDA_{MCD} and CQDA (QDA_C). The simulation study consists of different percentages of contamination in generated population groups. The empirical results of comparison show that the proposed robust discriminant rule performed better than other compared methods. The proposed RQDA is also applied in real dataset as well to understand the efficacy of the proposed method. All these investigations supports the use of RQDA_{S_n} for discriminant analysis in high dimensional datasets.

Chapter 6

Conclusion and Future Research Directions

Statistical analytical techniques are often used to extract information from the dataset. The major challenge of making a correct inference depends on the handling of outlying observations in the dataset. The outlier detection problem becomes more complex when the dimension of the data increases (curse of dimensionality). The exclusive use of a univariate outlier detection method to detect multivariate outliers is no longer excusable. The introductory chapter gives a brief idea about the multivariate statistical techniques and the importance of outlier detection in multivariate data. An extensive literature study of existing methods has been conducted in the introductory chapter on multivariate outliers detection and robust estimation of mean vector and covariance matrix.

Multivariate regression analysis is the statistical technique used to identify the relation between multivariate random variables. To propose an efficient robust multivariate regression analysis (Sajana and Sajesh 2018a), a multivariate robust estimation procedure called comedian method proposed by Sajesh and Srinivasan (2012) is adopted. The empirical study reveals that the comedian method showed less error performance than other methods existed in the literature for the estimation of multivariate robust regression. The technique is applied in the real-life dataset and it supports the results from simulated samples.

A bivariate robust dispersion estimator termed as S_nCov is developed out of

univariate robust S_n scale estimator proposed by (Sajesh and Srinivasan 2019). A robust correlation estimator is established corresponding to the proposed covariance estimator. Properties of the proposed estimator are demonstrated using theoretical and Mote Carlo methods. The efficiency of the proposed robust correlation is compared with some other popular robust correlation methods in terms of weighted MSE values and it showed better performance relative to others compared methods.

In the current days of advanced technology, the number of multivariate datasets that are being recorded every second is huge. The initial concern of a statistician is the cleansing of the available data and analyzing the data to reveal the mysteries contained in it. The most adaptive solution for minimizing the effect of outlying observation in the statistical analysis is the robustness of the statistical techniques. Well known methods for outlier detection and robust estimation like MCD, Kurtosis and OGK are highly affected by swamping and masking problems due to the presence of outlying observations in the data. Recent methods like SM method, RVMD, and MRCD are less affected by masking and swamping but have low breakdown values. According to reviews of available literature, the comedian method posses an optimum breakdown value and is less affected by swamping and masking issues. But still, there is space for improvement in terms of various factors.

This study proposes a robust technique for multivariate outlier detection by developing robust estimates for location vector and scatter matrix using S_n covariance proposed by Sajana and Sajesh (2019). The initial dispersion is not positive semi-definite (positive definite). An orthogonalization technique adopted from Maronna and Zamar (2002) is applied to make scatter matrix positive definite. Mahalanobis distance using these estimates is then used for multivariate outlier detection and robust estimation of scatter matrix and location vector. Simulation has been conducted by S_n method. The influence of masking and

swamping effects of multivariate outliers are evaluated using the parameters RSD and RFD. S_n method is compared with seven existing methods. The results show that the proposed S_n estimators can effectively overcome masking and swamping problems. The affine equivariance of the proposed S_n estimators are examined by generating correlated samples and the results are similar to that of uncorrelated samples. Simulation study also proved that the S_n posses optimum breakdown value. The S_n method is able to detect all the possible outliers presented in the dataset with minimum masking and swamping effect.

The main aim of the proposed S_n method is the development of robust multivariate techniques. The discriminant analysis is one of the most important multivariate data analysis techniques for classification. The S_n method approach of robust quadratic discriminant analysis has been established. The simulated population groups of different sample sizes and levels of contaminations are generated for calculating over all misclassification probabilities and the comparison are performed with quadratic discriminant analysis using classical, MCD and Comedian estimates. The proposed method has lower misclassification compared to the other methods for various population groups.

In conclusion, new robust alternative for covariance estimator is introduced and its multivariate extension towards outlier detection as well as robust estimation of mean vector and covariance matrix are studied in this thesis. The study starts with introducing relatively fast and efficient robust multivariate regression on the basis of comedian estimates. Further, the efficiencies and properties are examined through simulated samples. Later, an efficient and robust alternative for covariance estimators is developed based on robust scale estimator. The necessary properties are investigated by theoretical and empirical methods. Further, an efficient method for outlier detection and robust estimation of location vector and scatter matrix have been established. The proposed estimate, called S_n method has been developed based on robust Mahalanobis distance. Performance

of the method has been analyzed through simulation technique and tested in real datasets as well. Then, the application of proposed robust estimation method in robust discriminant analysis has been carried out. The efficiency of the proposed robust discriminant analysis is examined using simulated training dataset and validation datasets.

Multivariate statistical analysis rely on presuppositions. Suitable techniques are need to be undertaken to reduce the inconvenience that is made by the violation of these assumptions. It is relevant to extend the robustness in to the multivariate techniques that use linear models of continuous measurements. New method for outlier detection and robust estimations of multivariate location and scatter are considered in this thesis. Therefore, it is able to use outlier resistant estimation procedure to the data reduction techniques such as principal component analysis and factor analysis. This also can be applied in all ranges of design of experiments as well as MANOVA and related analysis techniques. The robust estimation of parameters that has been the functions of location vector and scatter matrix are used in regression analysis as well.

Barnett and Lewis (1994) have studied different methods of detecting outliers from continuous data. The multivariate data can also be categorical or mixed type in nature. In this case the application of the outlier detection and robust estimation can be adopted for robust analysis of similar type of data.

Appendix A

Tables of Chapter 2

Table A.1: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = 4$ and $q = 10$

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.324	1.101	1.066	1.074
Intercept	1.279	1.072	1.022	1.016
$\mathbf{\Sigma}_{diagonal}$	3.723	2.632	2.596	2.554
$\mathbf{\Sigma}_{off-diagonal}$	0.786	0.969	0.992	0.994
MLE:				
Slope	1.317	1.062	1.020	1.022
Intercept	1.272	1.048	0.984	0.995
$\mathbf{\Sigma}_{diagonal}$	3.686	2.368	2.258	2.094
$\mathbf{\Sigma}_{off-diagonal}$	0.782	0.944	0.977	0.967
MCD:				
Slope	5.219	1.609	1.251	1.182
Intercept	2.992	1.342	1.090	1.095
$\mathbf{\Sigma}_{diagonal}$	7.583	3.012	2.528	2.356
$\mathbf{\Sigma}_{off-diagonal}$	2.539	1.523	1.212	1.137
OGK:				
Slope	1.958	1.419	1.354	1.335
Intercept	1.770	1.297	1.156	1.212
$\mathbf{\Sigma}_{diagonal}$	7.676	6.571	8.323	11.632
$\mathbf{\Sigma}_{off-diagonal}$	0.755	1.029	1.059	1.052

Table A.2: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$ when the data contains 40% of vertical outliers

	n			
	50	200	500	1000
Comedian regression:				
Slope	12.487	6.554	8.777	13.157
Intercept	170.516	373.544	1602.672	5444.999
$\mathbf{\Sigma}_{diagonal}$	5304.519	16483.99	72856.47	250612.3
$\mathbf{\Sigma}_{off-diagonal}$	5516.615	16957.59	74829.62	2572711.1
MLE:				
Slope	31.204	26.305	26.876	25.952
Intercept	846.075	3358.234	8386.028	16783.810
$\mathbf{\Sigma}_{diagonal}$	24526.030	116970.600	301261.400	608892.0
$\mathbf{\Sigma}_{off-diagonal}$	25482.450	120633.600	310316.500	626954.0
MCD:				
Slope	77.393	50.714	50.593	47.986
Intercept	1818.105	9090.076	22660.577	45244.190
$\mathbf{\Sigma}_{diagonal}$	50644.300	185423.200	474748.800	962059.400
$\mathbf{\Sigma}_{off-diagonal}$	52398.640	193795.500	496380.500	1006114.300
OGK:				
Slope	51.700	30.970	30.826	29.763
Intercept	1010.802	4416.548	11286.298	22811.280
$\mathbf{\Sigma}_{diagonal}$	22468.810	122102.500	317716.600	644000.400
$\mathbf{\Sigma}_{off-diagonal}$	23621.880	127502.900	331643.200	672213.500

Table A.3: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = 6$ and $q = 4$ when the data contains 20% of vertical outliers

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.613	1.3602	1.308	1.337
Intercept	1.571	1.304	1.252	1.273
$\mathbf{\Sigma}_{diagonal}$	3.493	3.036	3.010	31.143
$\mathbf{\Sigma}_{off-diagonal}$	1.119	1.261	1.257	121.721
MLE:				
Slope	19.183	16.633	15.994	15.874
Intercept	188.919	744.910	1857.580	3715.516
$\mathbf{\Sigma}_{diagonal}$	8547.850	40742.983	105371.9	212954.3
$\mathbf{\Sigma}_{off-diagonal}$	8896.958	41909.972	108167.6	218371.50
MCD:				
Slope	3.316	1.605	1.396	1.383
Intercept	3.599	1.409	1.299	1.303
$\mathbf{\Sigma}_{diagonal}$	78.616	9.352	19.401	37.596
$\mathbf{\Sigma}_{off-diagonal}$	77.112	2.222	1.955	186.031
OGK:				
Slope	2.133	1.693	1.676	1.766
Intercept	1.839	1.472	1.518	1.623
$\mathbf{\Sigma}_{diagonal}$	4.999	4.546	5.639	778.42
$\mathbf{\Sigma}_{off-diagonal}$	1.047	1.368	1.501	165.740

Table A.4: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = 4$ and $q = 8$ when the data contains 20% of vertical outliers

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.511	1.350	1.312	1.317
Intercept	1.403	1.319	1.247	1.268
$\mathbf{\Sigma}_{diagonal}$	2.948	2.692	2.787	2.975
$\mathbf{\Sigma}_{off-diagonal}$	1.153	1.225	1.255	1.273
MLE:				
Slope	20.427	18.141	17.281	17.807
Intercept	210.978	839.864	2098.754	4191.323
$\mathbf{\Sigma}_{diagonal}$	12056.328	54645.416	136055.9	273853.9
$\mathbf{\Sigma}_{off-diagonal}$	12431.623	53360.474	139133.8	279965.6
MCD:				
Slope	9.486	1.792	1.395	1.367
Intercept	62.724	8.422	1.288	1.291
$\mathbf{\Sigma}_{diagonal}$	5848.072	562.044	17.395	3.176
$\mathbf{\Sigma}_{off-diagonal}$	5924.811	567.824	1.891	1.874
OGK:				
Slope	1.970	1.637	1.619	1.716
Intercept	1.662	1.507	1,478	1.574
$\mathbf{\Sigma}_{diagonal}$	4.148	4.096	5.006	6.895
$\mathbf{\Sigma}_{off-diagonal}$	1.114	1.277	1.361	1.473

Table A.5: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = q = 6$ when the data contains 40% of bad leverage points

	n			
	50	200	500	1000
Comedian regression:				
Slope	2.925	6.511	15.399	29.371
Intercept	2.377	2.301	2.653	2.799
$\mathbf{\Sigma}_{diagonal}$	7.882	25.653	68.981	138.443
$\mathbf{\Sigma}_{off-diagonal}$	1.186	2.247	4.829	8.849
MLE:				
Slope	3.182	6.998	15.173	28.843
Intercept	2.325	2.032	2.250	2.420
$\mathbf{\Sigma}_{diagonal}$	8.525	17.707	36.775	69.221
$\mathbf{\Sigma}_{off-diagonal}$	1.082	2.914	6.376	11.892
MCD:				
Slope	6.967	11.269	19.374	32.617
Intercept	11.637	16.036	21.838	26.879
$\mathbf{\Sigma}_{diagonal}$	22.341	100.072	267.437	565.172
$\mathbf{\Sigma}_{off-diagonal}$	0.828	0.901	1.152	1.567
OGK:				
Slope	4.831	8.577	16.855	30.973
Intercept	5.408	4.364	4.509	4.662
$\mathbf{\Sigma}_{diagonal}$	24.974	77.271	176.076	344.109
$\mathbf{\Sigma}_{off-diagonal}$	0.404	1.090	2.256	4.085

Table A.6: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = 6$ and $p = 4$ when the data contains 30% of bad leverage points

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.904	1.585	1.541	1.528
Intercept	1.768	1.531	1.490	1.497
$\mathbf{\Sigma}_{diagonal}$	4.067	1.419	3.378	3.553
$\mathbf{\Sigma}_{off-diagonal}$	1.246	3.497	1.454	1.461
MLE:				
Slope	2.599	5.599	12.036	22.836
Intercept	1.128	1.751	1.811	1.911
$\mathbf{\Sigma}_{diagonal}$	5.795	10.088	19.033	9.951
$\mathbf{\Sigma}_{off-diagonal}$	1.128	2.495	5.249	34.376
MCD:				
Slope	7.508	10.710	16.965	27.591
Intercept	9.203	9.179	8.329	7.734
$\mathbf{\Sigma}_{diagonal}$	17.458	57.840	117.627	213.090
$\mathbf{\Sigma}_{off-diagonal}$	1.485	2.082	3.635	5.605
OGK:				
Slope	2.328	3.378	7.510	15.562
Intercept	1.990	1.750	1.858	1.956
$\mathbf{\Sigma}_{diagonal}$	5.981	7.064	11.540	19.888
$\mathbf{\Sigma}_{off-diagonal}$	1.160	2.014	4.391	9.046

Table A.7: Finite sample comparison of Comedian, MLE, MCD and OGK estimators based on MSE for $p = 4$ and $p = 8$ when the data contains 30% of bad leverage points

	n			
	50	200	500	1000
Comedian regression:				
Slope	1.748	1.582	1.520	1.500
Intercept	1.665	1.493	1.486	1.492
$\mathbf{\Sigma}_{diagonal}$	3.280	3.020	3.121	3.317
$\mathbf{\Sigma}_{off-diagonal}$	1.302	1.392	1.428	1.477
MLE:				
Slope	6.184	19.748	47.197	92.957
Intercept	2.224	2.327	2.868	3.762
$\mathbf{\Sigma}_{diagonal}$	2.784	2.683	2.559	2.708
$\mathbf{\Sigma}_{off-diagonal}$	4.378	15.630	37.905	73.778
MCD:				
Slope	9.930	25.153	53.733	100.152
Intercept	7.048	9.617	10.766	10640
$\mathbf{\Sigma}_{diagonal}$	8.110	34.148	79.427	144.498
$\mathbf{\Sigma}_{off-diagonal}$	5.054	8.069	16.733	31.765
OGK:				
Slope	2,584	3.543	5.306	7.548
Intercept	1.942	1.692	1.748	1.801
$\mathbf{\Sigma}_{diagonal}$	4.188	4.166	5.167	7.555
$\mathbf{\Sigma}_{off-diagonal}$	1.676	3.247	5.460	8.766

Appendix B

R code for S_n method

```
SnCov=function(x){
p=ncol(x);n=nrow(x)
covSn=function(x,y){
n=length(x)
v=rep(0,n)
for(i in 1:n){
v[i]=((x[i]-x)*(y[i]-y))[-c(i)][order(((x[i]-x)*(y[i]-y))[-c(i)])][n%/%2]
}
return(sncov=v[order(v)][(n+1)%/%2])
}
V=matrix(0,p,p)
D=matrix(0,p,p)
for(j in 1:p){
for(i in 1:p){
if(i<=j){
V[i,j]=covSn(x[,i],x[,j])
D[i,i]=1/sqrt(V[i,i])
}
}}
V[lower.tri(V)]=t(V)[lower.tri(V)]
delta=D%*%V%*%t(D)
ei=eigen(delta)
Q=solve(D)%*%(ei$vectors)
z=t(solve(Q)%*%t(x))
gamma=matrix(0,p,p)
```

```
for(i in 1:p){
for(j in 1:p){
if(i==j){ gamma[i,j]=covSn(z[,i],z[,i]) }}}
I=cbind(colMedians(z))
scatter=Q%*%gamma%*%t(Q)
location=Q%*%I
RD=mahalanobis(x,center=location,cov=scatter)
cv1=qchisq(0.95,df=p)* ifelse(p<5,1,2.5)
cv2=qchisq(0.95,df=p)*median(RD)/qchisq(0.5,df=p)
cv=ifelse(p>=15,cv2,cv1)
wt=ifelse(RD>cv, 0, 1)
wtx=matrix(0,n,p)
k=1
for(i in 1:n){
if(wt[i]==1){
wtx[k,]=x[i,]
k=k+1
}}
wx=wtx[1:sum(wt),]
wcenter=colMeans(wx)
wcov=cov(wx)
result=list(weight=wt,RD=RD,cv=cv,wcenter=wcenter,wcov=wcov)
return(result)
}
```

Bibliography

Abdullah, M.B. 1990. On a robust correlation coefficient. *Journal of the Royal Statistical Society. Series D (The Statistician)* 39(4): 445–60.

Adebanji, A., M. Asamoah-Boaheng., and O. Osei-Tutu. 2016. Robustness of the quadratic discriminant function to correlated and uncorrelated normal training samples. *SpringerPlus* 5(102). doi:10.1186/s40064-016-1718-3.

Agostinelli, C, and L. Greco. 2019. Weighted likelihood estimation of multivariate location and scatter. *TEST* 28(3): 756–84.

Ahn, J., M.H. Lee, and J.A. Lee. 2019. Distance-based outlier detection for high dimension, low sample size data. *Journal of Applied Statistics* 46(1): 13–29. -

Ammeraal, L. 1992. *Programming principles in computer graphics*. Wiley, New York, USA.

Anderson, T.W. 2004. *An introduction to multivariate statistical analysis*. New York

Anscombe, F.J. 1973. Graphs in statistical analysis. *American Statistician* 27(1):17–21.

Atkinson, A.C. 1993. 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.

Atkinson, A.C. 1994. Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association* 89(428): 1329–39.

Bai, Z.D., X.R. Chen, B.Q. Miao, and C.R. Rao. 1990. Asymptotic theory of least distances estimate in multivariate linear models. *Statistics* 21(4): 503–19.

Barnett, V, and T. Lewis. 1994. *Outliers in statistical data*. 3rd Edition. New York: John Wiley and Sons.

Becker, C, and U. Gather. 1999. "The masking breakdown point of multivariate outlier identification rules." *Journal of the American Statistical Association* 94(447): 947–55.

Bebbington, A.C. 1978. "A method of bivariate trimming for robust estimation of the correlation coefficient." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 27(3): 221–26.

Billor, N., A.S. Hadi, and P.F. Velleman. 2000. BACON: Blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis* 34(3): 279–98.

Boudt, K., P.J. Rousseeuw, S. Vanduffel, and T. Verdonck. 2019. The minimum regularized covariance determinant estimator. *Statistics and Computing*. doi: 10.1007/s11222-019-09869-x.

Brown, B. M. 1983. Statistical uses of the spatial median. *Journal of the Royal Statistical Society, Series B* 45(1): 25-30.

Butler, R.W., P.L. Davies, and M. Jhun. 1993. Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics* 21(3): 1385–400.

Campbell, N.A. 1980. Robust procedures in multivariate analysis I: Robust covariance estimation. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29(3): 231–37.

Campbell, N. A. 1982, Robust procedures in multivariate analysis II: Robust canonical variate analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31(1): 1–8.

Campbell, N. A. 1989. Robust Bushfire mapping using NOAA AVHRR data: Technical Report, CSIRO, North Ryde, Australia.

Caroni, C, and P. Prescott. 1992. Sequential application of Wilks's multivariate outlier test. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41(2): 355–64.

Cerioli, A. 2010. Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association* 105(489): 147–56.

Chiang, L.H., R.J. Pell, and M.B. Seasholtz. 2003. Exploring process data with the use of robust outlier detection algorithms. *Journal of Process Control* 13(5): 437–49.

Croux, C., and C. Dehon. 2001. Robust linear discriminant analysis using S-Estimators. *The Canadian Journal of Statistics*, 29(3): 473–93.

Croux, C, and C. Dehon. 2010. Influence functions of the Spearman and Kendall correlation measures. *Statistical methods and Application* 19(4): 497–515.

Davies, L. 1992. The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics* 20(4): 1828–43.

Devlin, S. J, R. Gnanadesikan, and J. R. Kettenring. 1981. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association* 76(374): 354–62

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 P.J. Huber. 1982. The Notion of breakdown Point. In *A Festschrift for Erich L. Lehmann*. Bickel, P.J., K.A. Doksum, and J.L. Hodges Jr., eds, Berkeley: university of California.

Egan, W.J, and S.L. Morgan. 1998. Outlier detection in multivariate analytical chemical data. *Analytical Chemistry* 70(11): 2372–79.

Falk, M. 1997. On mad and comedians. *Annals of the Institute of Statistical Mathematics* 49(4): 615–44.

Fisher, R.A. 1936. The statistical utilization of multiple measurements. *Annals of Eugenics* 8: 376–86.

Gasko, M, and D.L. Donoho. 1982. Influential observations in data analysis. *American Statistical Association Proceedings of the Business and Economic Statistics Section*. 1: 104–9.

Gao, S., G. Li, and D. Wang. 2005. A new approach for detecting multivariate outliers. *Communications in Statistics-Theory and Methods* 34(8): 1857–65.

Gnanadesikan, R, and J.R. Kettenring. 1972. Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics* 28(1): 81–124.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* 11(1): 1–21.

Hadi, A.S. 1992. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)* 54(3): 761–71.

Hadi, A.S. 1994. A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society. Series B (Methodological)* 56(2): 393–96.

Hall, P, and A.H. Welsh. 1985. Limit theorems for the median deviation. *Annals of the Institute of Statistical Mathematics* 37(1): 27–36.

Hampel, F.R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69(346): 383–93.

Hampel, F.R. 1971. A general qualitative definition of robustness. *The Annals of Mathematical Statistics* 42(6): 1887–96.

Hampel, F.R. 1968. Contributions to the theory of robust estimation. PhD diss., Dept. Statistics, Univ.California, Berkeley.

Hampel, F.R., E. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. 1986. *Robust Statistics: The approach based on influence functions*. Wiley, New York.

Hawkins, D.M. 1994. The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis* 17(2): 197–210.

Hawkins, D.M. 1980. *Identification of Outliers*. Chapman and Hall, London.

Hodge, J.L. 1967. 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.

Hössjer, O., P.J. Rousseeuw, and C. Croux. 1992. Influence function and asymptotic normality of the repeated median slope estimator. Report 1992:2, Department of Mathematics, Uppsala University.

Huber, P.J. 1981. *Robust Statistics*. New York: John Wiley and Sons.

Huber, P.J. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35(1): 74–101.

Hubert, M., P.J. Rousseeuw, and K. Vakili. 2014. Shape bias of robust covariance estimators: an empirical study. *Statistical Papers* 55(1): 15–28.

Hubert, M., P.J. Rousseeuw, and T. Verdonck. 2012. A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics* 21(3): 618–37

Hubert, M., P.J. Rousseeuw, D. Vanpaemel and T. Verdonck. 2015. The DetS and DetMM estimators for multivariate location and scatter. *Computational Statistics and Data Analysis* 81: 64–75.

Hubert, M., and K. Van Driessen. 2004. Fast and robust discriminant analysis *Computational Statistics and Data Analysis* 45: 301–320.

Jackson, J.E, and G.S. Mudholkar. 1979. Control procedures for residuals associated with principal component analysis. *Technometrics* 21(3): 341–49.

Juan, J, and F.J. Prieto. 2001. Using angles to identify concentrated multivariate outliers. *Journal of the American Statistical Association* 43(3): 311–22.

Johnson, R.A, and D.W. Wichern. 1992. *Applied Multivariate Analysis*. Prentice-Hall of India Private Limited, New Delhi.

Koenker R, and S. Portnoy. 1990. M estimation of multivariate regressions. *Journal of the American Statistical Association*, 85(412):1060–1068.

Kim, M.G. (2000). Multivariate outliers and decompositions of Mahalanobis distance. *Communications in Statistics-Theory and Methods* 29(7): 1511–26.

Kirschstein, T., S. Liebscher, and C. Becker. 2013. Robust estimation of location and scatter by pruning the minimum spanning tree. *Journal of Multivariate Analysis* 120: 173–84.

Kosinski, A.S. (1998). A procedure for the detection of multivariate outliers. *Computational Statistics and Data Analysis* 29(2): 145–61.

Lakshmi, R. and T.A. Sajesh (2018). Robust quadratic discriminant analysis using Kurtosis method. *Journal of Computer and Mathematical Sciences* 9(12): 1907–14.

Lee, J. 1992. Relationships between properties of Pulp-Fibre and paper, PhD diss., U. Toronto, Faculty of Forestry.

Liebscher, S., T. Kirschstein, and C. Becker. 2012. RDELA-a Delaunay-triangulation-based, location and covariance estimator with high breakdown point. *Statistics and Computing* 23(6): 677–88.

Leys, C., O. Klein, Y. Dominicy, and C. Ley. 2018. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology* 74: 150–6.

Li, Zh.V, G.L. Shevlyakov and V.I. Shin. 2006. Robust estimation of correlation coefficient for ε-contaminated bivariate normal distributions. *Automation and Remote Control* 67(12):1940–57.

Lim, Yai-Fung., Sharipah-Soaad Syed-Yahaya., and Hazlina Ali. 2017. Robust linear discriminant analysis with distance based estimators. Proceedings of the 13th IMT-GT International Conference on Mathematics, Statistics and their Applications (ICMSA2017). doi: org/10.1063/1.5012246

Lopuhaä, H.P. and P.J. Rousseeuw. 1991. Breakdown Points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics* 19(1): 229–48.

Lopuhaä, H.P. 1989. On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics* 17(4): 1662–83.

Ma, Y, and M.G. Genton. 2001. Highly robust estimation of dispersion matrices. *Journal of Multivariate Analysis* 78(1): 11–36.

Mahalanobis, P. C. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences* 2: 49–55.

Maronna, R.A. 1976. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics* 4(1): 51–67.

Maronna, R. A, and V. J. Yohai. 1995. The behaviour of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association* 90(429): 330–41

Maronna, R A, and V.J. Yohai. 1997. Robust estimation in simultaneous equations Models. *Journal of Statistical Planning and Inference* 57(2): 233–44.

Maronna, R.A, and R.H. Zamar. 2002. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* 44(4): 307–17.

Ollila, E., H. Oja, and T.P. Hettmansperger. 2002. Estimates of regression coefficients based on sign covariance matrix. *Journal of the Royal Statistical Society, Statistical Methodology: Series B* 64(3): 447–66.

Oyeyemi, G.M. and Ipinyomi, R.A. 2010. A robust method of estimating covariance matrix in multivariate data analysis. *African Journal of Mathematics and Computer Science Research* 3(1): 1–18.

Pan, J.X., W.K. Fung, and K.T. Fang. 2000. Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference* 83(1): 153–67.

Peña, D, and F.J. Prieto. 2001. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics* 43(3): 286–310.

Peña, D, and F.J. Prieto. 2007. Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics* 16(1): 228–54.

Preparata, F. P, and M.I. Shamos. 1985. *Computational geometry: An introduction*. New York: Springer.

Pyke, R. 1965. Spacings. *Journal of the Royal Statistical Society. Series B (Methodological)* 27(3): 395–449.

Rao, C.R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhya:The Indian Journal of Statistics, Series A (1961-2002)* 26(4): 329–58.

Rocke, D.M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of Statistics* 24(3): 1327–45.

Rocke, D.M, and D.L. Woodruff. 1993. Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica* 47(1): 27–42.

Rocke, D.M, and D.L. Woodruff. 1996. Identification of outliers in multivariate data. *Journal of the American Statistical Association* 91(435): 1047–61.

Ro, K., C. Zou, Z. Wang, and G. Yin. 2015. Outlier detection for high-dimensional data. *Biometrika* 102(3): 589–99.

Rousseeuw, P. J. 1985. Multivariate estimation with high breakdown point. *In Mathematical Statistics and Applications* B (W. Grossmann, G. Pflug, I. Vincze and W. Werz, eds.) 283–297. Reidel, Dordrecht.

Rousseeuw, P.J, and C. Croux. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88(424): 1273–83.

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

Rousseeuw, P.J, and K. Van Driessen 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41(3): 212–23.

Rousseeuw, P.J, and B.C. Van Zomeren. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85(411): 633–39.

Rousseeuw, P.J., S. Van Aelst, K. Van Driessen, and J. Agulló. 2004. Robust multivariate regression. *Technometrics* 46(3): 293–305.

Rousseeuw, P.J, and V.J. Yohai. 1984. Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics* 26: 256–72.

Sajana, O.K, and T.A. Sajesh. 2018. Empirical robust multivariate regression parameter estimation using median approach. *International Journal of Scientific Research in Mathematical and Statistical Sciences.* 5(5): 65–71.

Sajana, O. K, and T. A. Sajesh. 2018. Detection of multidimensional outlier using multivariate spatial median. *Journal of Computer and Mathematical Sciences* 9(12): 1875–81.

Sajana, O.K, and T.A. Sajesh. 2019. S_n covariance. *Communications in Statistics-Theory and methods.*doi: 10.1080/03610926.2019.1628275.

Sajana, O.K, and T.A. Sajesh. 2020. Multidimensional outlier detection and robust estimation using S_n Covariance. *Communications in Statistics-Simulation and Computation.*doi: 10.1080/03610918.2020.1725820

Sajesh, T.A, and M.R. Srinivasan. 2012. Outlier detection for high dimensional data using the comedian approach. *Journal of Statistical Computation and Simulation* 82(5): 745–57.

Sajesh T.A, and M.R. Srinivasan. 2013. An overview of multiple outliers in multidimensional data. *Sri Lankan Journal of Applied statistics* 14(2): 87–120.

Sajesh T.A, and M.R. Srinivasan. 2019. Robust quadratic discriminant rule using Comedian. *Research and Review: Journal of Statistics* 8(2): 41–47.

Salibián-Barrera, M., S. Van Aelst, and G. Willems. 2006. Principal components analysis based on multivariate MM estimators with fast and robust bootstrap. *Journal of the American Statistical Association* 101(475): 1198–211.

Salibián-Barrera, M, and V.J. Yohai. 2006. A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics* 15(2): 414–27.

Shevlyakov, G.L. 1997. On robust estimation of a correlation coefficient. *Journal of Mathematical Sciences* 83(3):434–38.

Shevlyakov, G.L, and P. Smirnov. 2011. Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics* 40: 147–56.

Siegel, A.F. 1982. Robust regression using repeated medians. *Biometrika* 69(1): 242–44.

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

Tukey, J.W. 1960. A survey of sampling from contaminated distributions. *Contributions to probability and statistics: Essays in honor of Harold Hotelling* 69: 448–85.

Tukey, J.W. 1977. *Exploratory data analysis*. Reading: Massachusetts: Addison-Wesley.

van Capelleveen, G., M. Poel, R.M. Mueller, D. Thornton, and J. van Hillegersberg. 2016. Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *International Journal of Accounting Information Systems* 21: 18–31

Viljoen, H, and J.H. Venter. 2002. Identifying multivariate discordant observations: A computer-intensive approach. *Computational Statistics and Data Analysis* 40(1): 159–72.

Wang, M.Y, and C.E. Zwilling. 2015. Multivariate computing and robust estimating for outlier and novelty in data and imaging sciences. In *Advances in Bioengineering*, P.A. Serra, 317–36.

Wasi Haider Butt., M. Usman Akram, Shoab A. Khan, and Muhammad Younus Javed. 2014. Covert network analysis for key player detection and event prediction using a hybrid classifier. *The Scientific World Journal*. doi: 10.1155/2014/615431

Wilcox, R.R. 2017. *Introduction to Robust Estimation and Hypothesis Testing*. 4th. ed. Amsterdam: Elsevier

Wilks, S.S. 1963. Multivariate statistical outliers. *Sankhya: The Indian Journal of Statistics. Series A* 25(4): 407–26.

Woodruff, D.L, and D.M. Rocke. 1993. Heuristic search algorithms for the minimum volume ellipsoid. *Journal of Computational and Graphical Statistics* 2(1): 69–95.

Woodruff, D.L, and D.M. Rocke. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association* 89(427): 888–96.

Yang, X., Z. Wang, and X. Zi. 2018. Thresholding-based outlier detection for high-dimensional data. *Journal of Statistical Computation and Simulation* 88(11): 2170–84.

Yu, S, Z. Cao., and X. Jiang. 2017. Robust linear discriminant analysis with a Laplacian assumption on projection distribution. 2017 IEEE International

Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, 2567–71.

Zhao, H., Z. Wang., and F. Nie. 2018. A new formulation of linear discriminant analysis for robust dimensionality reduction. IEEE Transactions on Knowledge and Data Engineering, 31(4): 629–40. doi:10.1109/tkde.2018.2842023

Zani, S., M. Riani, and A. Corbellini. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data analysis* 28(3): 257–70.

List of the Published Works by the Researcher

1. **Published Papers**

   (a) Sajana, O.K, and T.A. Sajesh. 2018a. Empirical robust multivariate regression parameter estimation using median approach. *International Journal of Scientific Research in Mathematical and Statistical Sciences*. 5(5): 65–71. doi: 10.26438/ijcse/v5i5.6571.

   (b) Sajana, O. K, and T. A. Sajesh. 2018b. Detection of multidimensional outlier using multivariate spatial median. *Journal of Computer and Mathematical Sciences* 9(12): 1875–81. doi:http://dx.doi.org/10.29055/jcms/934.

   (c) Sajana, O.K, and T.A. Sajesh. 2019. S_n covariance. *Communications in Statistics-Theory and methods*. doi: 10.1080/03610926.2019.1628275.

   (d) Sajana, O.K, and T.A. Sajesh. 2020. Multidimensional outlier detection and robust estimation using S_n Covariance. *Communications in Statistics-Simulation and Computation*. doi:10.1080/03610918.2020.1725820.

   (e) Sajana, O.K, and T.A. Sajesh. Robust Quadratic Discriminant Analysis Using S_n covariance. Communicated for publication in *Communications in Statistics-Simulation and Computation*.

2. **Presentations in Conferences/Seminars**

   (a) Fast and robust Multivariate regression, *National Seminar on Applied Statistical Methodology with Special Emphasis on Time Series Analysis* conducted by Department of Statistics Nirmala College, Muvattupuzha, Kerala, February 12-13, 2016.

   (b) Robust Canonical Correlation, *Second International Conference on Statistics for Twenty-First Century* conducted by Department of Statis-

tics, Kerala university, Trivandrum, December 21-23, 2016.

(c) Robust Regression-An introduction and Comparison, *National Seminar on Statistical Techniques in Applied Areas* conducted by Department of Statistics, St Thomas College (Autonomous), Thrissur, Kerala, February 27-28, 2017.

(d) Robust Estimation using L1-Median and its Properties, *International Conference on Theory and Applications of Statistics and Information Sciences* conducted by Department of Statistics Bharathiar University, Coimbatore, January 5-7, 2018.

(e) Multivariate Robust Estimation of Location and Covariance and its Empirically Solved Properties, *International Conference on Changing Paradigm and Emerging Challenges in Statistical Sciences* conducted by Department of Statistics Pondicherry University, January 29-31, 2018.

(f) A Comedian Approach to Multivariate Regression, *National Seminar on Innovative Approaches in Statistics* conducted by Department of Statistics, St Thomas College (Autonomous), Thrissur, Kerala, February 15-17, 2018.

(g) Detection of Multidimensional Outlier Using Multivariate Spatial Median, *National Conference on Changing Paradigms in Teaching and Research of Statistics* conducted by Department of Statistics , Kristu Jayanti College (Autonomous) , Bangaluru, December 14, 2018.