OUTLIER DETECTION FOR SKEWED DISTRIBUTION: BIVARIATE CASE

BY

Umm-E-Furwa

Registration No: PIDE2015FMPHILETS09

MPhil Econometrics

A Dissertation Submitted to the Pakistan Institute of Development Economics, Islamabad, in partial fulfillment of the requirements of the Degree of Master of Philosophy in Econometrics

Supervised by

Dr. Atiq-ur-Rehman

Assistant Professor

Department of Econometrics and Statistics

Pakistan Institute of Development Economics
Islamabad, Pakistan

December, 2017

OUTLIER DETECTION FOR SKEWED DISTRIBUTION: BIVARIATE CASE

BY

Umm-E-Furwa

Registration No: PIDE2015FMPHILETS09

MPhil Econometrics

Supervised by

Dr. Atiq-ur-Rehman

Assistant Professor

Department of Econometrics and Statistics

Pakistan Institute of Development Economics
Islamabad, Pakistan

December, 2017

Pakistan Institute of Development Economics

CERTIFICATE

This is to certify that this thesis entitled: **"Outlier Detection for Skewed Distribution: Bivariate Case"** submitted by Ms. Umm-e-Furwa is accepted in its present form by the Department of Econometrics and Statistics, Pakistan Institute of Development Economics (PIDE), Islamabad as satisfying the requirements for partial fulfillment of the degree in **Master of Philosophy in Econometrics**.

Supervisor:

Dr. Atiq-ur-Rehman
Assistant Professor
PIDE, Islamabad

External Examiner:

Dr. Zahid Asghar
Associate Professor
Quaid-e-Azam University,
Islamabad

Head,
Department of Econometrics and Statistics

Dr. Amena Urooj


بسم الله الرحمن الرحيم

Disclaimer

This document represents part of the author's MPhil study program at Pakistan Institute of Development Economics. The views stated therein are those of the author herself and the work has been completed in a scheduled time.

Dedicated to my beloved father MUAHMMAD ASLAM and mother NAHEED and my Family

ACKNOWLEDGEMENT

All commend to ***ALLAH*** whose blessing and magnificence prosper my thoughts, and flourished my aims by giving me brilliant teachers, supporting parents and in-laws, exceptional friends and for giving me potential to completing this thesis. I would like to state here my supreme gratitude to the people who have assisted and supported me during my research.

Special thanks goes to my supervisor, ***Dr. Atiq ur Rehman***, Pakistan institute of Development Economic Islamabad, for his supervision and regular support. I am thankful to his ever encouraging guidance, keen attention, his precious assistance through beneficial remarks and suggestions during the research works have contribute to the accomplishment of this study.

Special thanks of mine goes to my friends ***Iqra Mazhar, Muneeza Maqbool*** and ***Sahar Arshad,*** who helped me in completing the thesis. Their association with their faithfulness has prompted and encouraged my intellect to maturity which is required for this work and I wish to continue our healthy friendship in the future.

Before I finish, my warmth gratitude goes to my beloved ***Parents*** and my ***loving father in-law Muhammad Naveed*** and ***mother in-law Shamim,*** my siblings and my sisters in law for their never-ending love, prayers and back-up, for their marvelous assistance and support both morally and financially towards the finishing of this study. To those who indirectly supported in this thesis, your concern means a lot to me. Thank you very much to all of you.

UMM-E-FURWA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

Acronyms	Description
AREA	Area of Fence
IQR_U	Upper Interquartile Range
IQR_L	Lower Interquartile Range
IRLS	Iteratively Reweighted Least Square
LCV_h	Lower Critical Value for Horizontal Distance
LCV_v	Lower Critical Value for Vertical Distance
MCD	Minimum Covariance Determinant
MD	Mahalanobis Distance
OD	Outlier Detection
RD	Robust Distance
SSS	Split Sample Skewness
SSSBB	Split Sample Skewness Based Boxplot
UCV_h	Upper Critical Value for Horizontal Distance
UCV_v	Upper Critical Value for Vertical Distance

ABSTRACT

Most of real data contains observations that might not be in the conformity of the rest of the data set. These observations are known to be outliers and might be caused by the personal mistake/error or due to natural variation. It is important to detect the outliers in the data set as outliers might have positive or negative effect on the regression analysis, forecasting results and ANOVA etc. Outliers are influential tools to classify the most remarkable events of the world in cross sectional data and generally important events can be chosen by detecting outliers in time series data sets. Numerous outlier detection techniques have been discussed in the literature for the detection of outliers in univariate, bivariate and multivariate data set. Most of these techniques work well when the data is normal but they give misleading results for the skewed data. There are various techniques to detect outliers in skewed data for univariate case but when we have more than one variable, there are very limited techniques as we consider the case of multivariate skewed data. As, multivariate data has many practical uses in real life and to find the relationship between the sets of variables, it's important to detect outliers in the multivariate case. Adil (2011) proposed a technique namely SSSBB to detect the outlier in the univariate skewed data only and proved that SSSBB performance is better than the existing ones. In this study, we have extended SSSBB for the bivariate case and compared the result with the robust Mahalanobis distance technique considering various types of distributions. This study uses Monte Carlo Simulations for comparison purposes of SSSBB and Mahalanobis distance. The study considered the normal distribution, chi-square, gamma and beta distributions and different sample sizes are taken, to evaluate the performance of SSSBB for bivariate data and the study found that SSSBB performs well as compared to Mahalanobis distance, in all the

cases considered in the thesis. On the basis of ratio of outlier detected and the area of fence, the results show that SSSBB is a better method for normal as well as skewed data sets because SSSBB detects the possible outliers in the specified area of fence.

CHAPTER: 1

1. INTRODUCTION

The observations that deviate markedly from rest of the observations in the sample and appear to be suspicious to the observer are called outliers. Outlier detection had been a serious issue for long time and is one of the earliest statistical interests. Since almost all data sets contain outliers of varying percentages therefore, it continues to be the most important issue. Sometimes outliers can distort the results, at other times their effect is unobservable. Therefore, many theories were generated related to it, like whether to keep these observations or to delete them from the data. Hodge and Austin (2004) discussed that outliers can be caused by many reasons like due to personal mistake, instrumental error or by natural deviations in the population, fraudulent behavior, changes in behavior of systems or faults in the systems. Many times the presence of outlier causes many difficulties like it can lead to the misspecification of the model, biased parameter estimation and bad forecasting in the estimation results. Osborne and Overbay (2004) illustrated that due to the presence of outliers the parameter estimation is highly influenced because it leads to the increase in the variance of the error and decreases in the power of the test. The presence of outlier in error term causes decrease in their normality in univariate and in the case of multivariate, causes decrease in the sphericity and multivariate normality. This leads to committing type 1 and type 2 errors.

According to Acuna et al (2004) different methods have been proposed for the detection of outliers therefore; the choice of these methods depends on the type of

outliers and the type of the data distribution. Some of the commonly used methods are Graphical technique, Grubbs test, 2SD and 3SD methods, Z-score and modified Z-score test and Dixon test. The details of the methods are discussed in chapter 2. A serious problem of the existing techniques is that they are applicable to symmetric distribution and fails to work in asymmetric distributions. For asymmetric distribution, the values of the variable occur at irregular frequencies with mean, median and mode at different points thus exhibiting skewness. The graph can be right skewed or left skewed depending on the type of data, but for the normal distribution it shows symmetry and a bell shape. Mostly, the outlier detection methods assume normality on contrary most of the real data exhibit non-normal behavior. Therefore, there is a need for a method that works well in many kinds of distributions equally. Tukey's method (1969) is designed keeping in view the skewed distribution; however, the method gives misleading results when the data is more skewed. Hubert and Vandervieren (2008) proposed a method for outlier detection in skewed distribution named as adjusted boxplot; it uses the 'med couple' which is a robust measure of skewness. The Median Absolute Deviation (MAD) is the basic robust technique; it is highly insensitive by the presence of extreme values in the data set but this method loses power when the skewness in data is moderate. Adil (2011) proposed a technique termed as "split sample skewness based boxplot" (SSSBB) that works equally well for symmetric and the asymmetric distributions. This technique provides greater accuracy in the detection of outliers when there is skewed data. Adil (2011) shows that the performance of SSSBB method is better than the existing techniques including Tukey's method, Adjusted boxplot, Hubert and Vandervian and Kimber, by comparing

their constructed fences with the true lower and upper boundaries around the central 95 percent of the distribution. SSSBB technique is applicable on a univariate data and by comparing the results; the study found that SSSBB technique to be better than the existing ones. The SSSBB method calculations are analytical and easy to understand. However, SSSBB technique is only applicable on univariate data. SSSBB method can be extended to bivariate and multivariate data and that is the aim of this thesis. Therefore, this thesis aims to extend the SSSBB technique developed by Adil for bivariate data and to compare its performance with the existing multivariate outliers detecting techniques.

1.1: OBJECTIVES OF THE STUDY

The objectives of the study are as follows:

i) To extend SSSBB for bivariate data.

ii) To compare the results of SSSBB in bivariate case with the robust Mahalanobis distance method for different distributions.

iii) To test the newly designed technique SSSBB for real data.

1.2: SIGNIFICANCE OF THE STUDY

Outliers cause many problems in the estimation results and it is important to handle them properly. Most of the methods proposed earlier are for univariate and bivariate data assuming symmetric distribution but they fail to give appropriate results when data follows asymmetric distribution. In this thesis a method is proposed for bivariate data that works equally well in symmetric as well as asymmetric distributions.

1.3: RESEARCH GAP

Most of the methods proposed for outlier detection considered normal distributions and less attention has been given for outlier detection in skewed distribution. Few researchers discussed about the outlier detection in univariate data for skewed distribution but for bivariate data no such work is done considering skewed distribution. In order to fill this gap a method has been proposed for bivariate data considering skewed distributions.

1.4: ORGANIZATION OF THE STUDY

In chapter 2, the study provides the literature review of the existence of outliers in the real data due to natural effects or sometime due to errors. The positive and the negative affect of outliers, different outlier detection techniques for univariate, bivariate and multivariate data will be discussed. Chapter 3 includes the methodology of the study. In this chapter we will convert SSSBB to the bivariate case. The performance of the test will be checked for the normal as well as the skewed distribution and will compare its results with a renowned technique "robust Mahalanobis distance" by using the algorithm of minimum covariance determinant method. Chapter 4 will be the interpretation of the results obtained in previous chapter by SSSBB in bivariate case for symmetric and skewed distributions. Furthermore, both the techniques are applied on Pakistan's Stock Exchange data and on the measures of interest rate. Chapter 5 is the conclusion/summary of the study and future work related to the study.

CHAPTER: 2

2. LITERATURE REVIEW

2.1: OUTLINE

The concern of detecting outliers is a serious issue to be considered in the studies because outlier causes serious problems in estimating different economic theories and give misleading results. In this regard, different outlier detection methods have been proposed in the literature, divided into three parts i.e. studies considering univariate data, studies with bivariate data and methods considering multivariate data, for symmetric and skewed distributions.

2.2: DEFINITON OF OUTLIER

Ordinarily we have the data set that contains discordant observations which look different from the other points but are present in the data due to some sort of connection **(**Edgeworth, 1887; cited by Beckman and Cook, 1983). Dixon (1950) defined outlier as "dubious in the eyes of the researcher". Generally, "objective" methods to deal with the problem of outliers would be employed only after the identification of outliers through a visual inspection of the data set (Grubbs, 1969; cited by Beckman and Cook, 1983).Weiner (1976) defined outliers as contaminant and believed that they have a disproportionate influence on the data. Some observations deviate remarkably from the rest of the observations as to create suspicions that they were created by a different

mechanism (Hawkins, 1980). Contaminants are the points which lie far outside the typically expected variation that can be sometimes noted or unnoted by the investigator (Barnett, 1984). Inconsistent observations that differ from the remaining data set are named as outlier (Iglewicz and Hoaglin, 1994). As this study is dealing with outliers therefore discordant observations, contaminants and inconsistent observations come under the definition of outliers.

2.3: HISTORY OF OUTLIER

Identification of outliers in the data analysis sets back to 18th century. Bernoulli (1777) pointed the deletion of outliers about 250 years ago, but it doesn't seem to be a proper way out to the problem of outliers. The first statistical technique to the problem of outliers was developed by Beckman and cook in 1850. Pierce (1852) , Chauvenet (1863) ,Wright (1884) and Cousineau and Chartier (2010) were in the favor of deleting the observations which were far away from the data but Bessel and baeuer (1838), Legendre (1852) were not in the favor of deleting the outliers. Bendre and Kale (1987), Davies and Gather (1993), Iglewicz and Hoaglin (1994) and Barnett and Lewis (1994) have conducted a number of studies to handle the issues of outliers. The viewpoint of Cousineau and Chartier (2010) opinion was that the as outliers were the outcome of some spurious action, hence they should be removed. Therefore, whether to delete or keep the outlier in the data is still a controversy today as it was 250 years ago.

2.4: CAUSES OF OUTLIERS AND HOW TO DEAL WITH THEM?

Anscombe (1960) has categorized causes of outliers into two ways; outlier may emerge due to some mistake/error and outlier may be caused as a result of natural variability. Ludbrook (2008) discussed many reasons of outlier's existence and the methods to handle them properly. Osborne and Overbay (2009) discussed the possible causes of outlier present in the data and how to deal with them in the following way:

- Outlier due to error in the data:

Outliers may arise due to human error such as during data collection, recording or entering the data. Errors of this nature can be corrected by referring to the original document and entering the correct values. Therefore, if there is appropriate information available, recalculation is a way to save the important data and exclude an obvious outlier. If outliers of this type cannot be modified, they should be eliminated as they do not denote valid population data points.

- Outliers from sampling error:

Outliers may be caused by sampling. If some data is taken from different population, then it may lead to generation of outliers. In this case, these points must be deleted as they are not representing the actual population.

- Outliers from standardization failure:

Outliers might appear by the research methodology, probably if something unusual happened during a specific subject experience. If such a situation occurs, researcher can delete such data if he is not interested in studying that particular type.

- Outliers from faulty distributional assumptions:

If wrong assumptions are taken about the distribution of the data, it can lead to the occurrence of suspected outliers in the data. May be the researcher assumed different data structure originally and it comes out to be different. It depends on the aim of the research whether to keep these extreme values or not.

2.4.1: OUTLIER DUE TO NATURAL DEVIATION

Natural variation is the evident matter which cannot be unnoticed and it is an observable outlier. For example, Zaineb Bibi the tallest woman in Pakistan, whose height is 7’3”, is an outlier as she is different from rest of the women population of Pakistan. But as it is the natural variation so this outlier cannot be ignored.

2.5: EFFECTS OF OUTLIERS

Outliers may have a positive or a negative effect on the data depending on the type of analysis. As in some cases we can remove the outlier if it doesn’t affect the results but sometimes outliers cannot be removed because they have some importance in the data and give some interesting information. Barnett (1978) discussed the famous case of

Hadlum vs Hadlum held in 1949, which is statistically an interesting case because of outlier. Normal gestation period is 280 days, but Mrs. Hadlum gave birth to her child after 349 days, so it is a natural outlier which cannot be disregarded. Outlier can have negative effect on the data if it comes by mistake. So, if outlier comes by mistake in the data, it provides misleading results.

Now we will see the positive and the negative effects of outliers.

2.5.1: POSITIVE EFFECTS OF OUTLIER ON DATA

In cross-sectional data, outliers expose interesting facts about the data which proves to be helpful for the researcher. Outlier appears to be different from the rest of the points thus; the researcher tries to find out the genuine cause of its appearance.

2.5.2: NEGATIVE EFFECTS OF OUTLIER ON DATA

Outlier seems to be suspicious to the researcher and if outlier is not detected at the right time, it causes error in the estimation of the parameter and the analysis of the data. Outlier significantly affects the estimation results, because outliers can cause increase in the standard error and decrease in the power of the test. The presence of outliers in errors results in the decrease of error normality in univariate case and sphericity and multivariate normality is affected in case of multivariate, thus altering the chances of producing the two types of errors, type I and type II errors. Osborne and Overbay (2004) illustrates the problem caused by outliers in the regression estimates being distorted by the outliers.

Hence, it is important for a researcher to identify the outliers in the start and take measures to avoid the problem in the end of the estimation.

2.6: IMPORTANCE OF DETECTING OUTLIER

Detection of outlier plays a significant role in modeling, inference and data handling because outlier may cause model misspecification, biased parameter estimation and bad forecasting (Tsay, Pena and Pankratz, 2000 and Fuller, 1987). Iglewicz and Hoaglin (1996) suggested that it's important to review the data for outliers as they can provide valuable information related to the data under consideration. The recognition of outliers may lead to the detection of unpredicted information in many fields such as credit card and calling card deceit, criminal activities, and cybercrime (Mansur and Sap, 2005). Outlier has an important role in data mining as the researchers interested in data mining have to aspect the problem of outliers that might arise from the real data generating process (DGP). Outliers are possibly to exist even in a high quality data set and very rare economic data sets encounter the benchmark of high quality (Zaman, Rousseeuw and Orhan, 2001).

2.7: METHODS FOR OUTLIER DETECTION IN UNIVARIATE DATA

Generally there are many methods for the outlier detection in the one-dimensional data set and they have some pros and cons. Outlier can also be identified by the graphical method and it is the simplest way to detect outlier.

Pierce (1852) was the first who proposed the criterion for the rejection of outlier. Stone (1867) gave another criterion, said that, "a person can commit, on average, one mistake in the making and registering of m (modulus of carelessness) observation of a given class", so the observations whose deviations have greater probabilities should be rejected. Chauvenet (1876) gave the criterion for the rejection of one observation. point of view was that, on the average a mistake occurs once in 2n observations, here n showed the number of observations in the sample under consideration.

The practice of deleting the outliers prevailed till 20^{th} century, but as outlier couldn't always be deleted so there should be some method that didn't involve the deletion of the outliers. Irwin (1925) proposed a statistic based on the fact that if the observations were arranged on the basis of magnitude, then by taking the difference between the pth and the (p+1)th observation, the frequency distribution could be obtained. So, it might be decided whether the extreme observations were from the same population or different. Tippet (1925) did work on the possibility of using the range to determine whether the outlying observation of a sample should be rejected or kept. McKay (1935) derived the approximate probability distribution for the extreme observation assuming a probability distribution of unity. For m greater than 3, the probability function was derived to evaluate the approximate values for the extreme values. Nair (1948) derived the exact values of probability function and matched them with the approximated values. Thompson (1935) assumed the random sample for normal distribution and proposed probability distribution function. The observations whose value was greater than student's t value, considered as outlier. Thompson's criterion differed

from the previous techniques in a way it didn't require σ to be known and it referred to arbitrary observations. Walsh (1950) proposed a non-parametric test to check whether outlying observations were from the same population or different. Grubbs (1950) introduced the method for the outlier detection in univariate normal distribution, considering the sample size larger than 3. Grubbs used mean and standard deviation in the technique and the largest absolute value was considered as outlier. However, Grubb's technique didn't discuss about the problem of extreme observation at either end. 2SD and 3SD method involved the construction of intervals $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$; here μ represents the sample mean and σ the standard deviation of the model. The problem in both the techniques was that they perform well in normal distribution and fails in skewed distribution. Dixon (1963) proposed technique based on "sub range ratio" for the data transformed in any order. The technique was applicable on normal small data sets and detected small numbers of outliers. If a value was observed as outlier by Dixon test it was checked in critical value table that whether it is an outlier or inlier. The main problem in Dixon's test was that if one value detected as outlier, then the test could not be applied for the same remaining data again. Tukey (1969) proposed boxplot based on first and third quartile and interquartile range. These boxplots provided better data summaries as compared to other methods, but the problem raised when the data was more skewed as boxplot then gives misleading results. Iglewicz and Hoaglin (1993) proposed test statistics based on median and median of the absolute deviation for univariate data. Carling (1998) proposed a technique for univariate skewed data that was based on median rule and interquartile range for the detection of outliers. Vanderviere and Huber

(2004) presented an adjusted boxplot using medcouple (MC) that was a robust measure of skewness for a skewed distribution, the observations which lie outside the interval were considered as outliers.

Adil (2011) proposed SSSBB for outlier detection in both symmetric and non-symmetric data by calculating skewness by split sample skewness (SSS). Calculating the Q_{1L} and IQR_L makes the lower critical value and Q_{3R} and IQR_R makes the upper critical value. The values which lie outside the interval of lower and upper critical values were labeled as outliers.

2.8: METHODS FOR OUTLIER DETECTION IN MULTIVARAITE DATA

Visual inspection doesn't work in multivariate scenario, as it becomes more difficult to detect outlier with the increase in the outlier numbers and the dimension of the data, because outliers do not stick out on the end and can grow in any number of directions. Multivariate data comprise more than one variable therefore, the outlier detection technique is used by getting more than one variable to interact with one another and identify the unusual observations. It is not necessary that the observations which are outliers in multivariate case would be outlier in univariate subset as well. Robust measures in the outlier detection techniques in multivariate data improve the performance of the technique. To have a successful method for multivariate outlier detection, it should be highly sensitive to the extreme values; the capacity to detect genuine outliers and highly specific; the ability that should not mistakenly detect regular observations as outlier. Gnanadesikan and Kettenring (1972) quote "it would be fruitless to search for a

truly omnibus outlier detection procedure." The multivariate location and shape is problematic, because most of the methods would break down if the part of outliers is greater than $1/(p+1)$, where p represents the dimension of the dataset (Maronna, 1976; Donoho, 1982; Stahel, 1981).

Wilks (1963) worked out on the problem of detecting outliers in a multivariate normal distribution with unknown parameters. Although, the desired results were not obtained but the test criterion for the single and pair of outliers was generated, but the problem arises when there were more than two outliers.

Outlier detection in multivariate data involved the methods of robust distances, these methods used robust estimators to calculate the mean vector and the covariance matrix and the Mahalanobis distance would be calculated for every observation using these robust estimates, the distance for the observation which exceeds the critical values were considered as outliers.

To calculate the robust distance for outlier detection, Campbell (1980) used M-estimator to find the robust mean vector, the covariance matrix and calculated the Mahalanobis distance using these estimates. The points distant from the bulk of the data were considered as outliers. In higher dimension, M-estimator was greatly affected by the outlying observations and the estimator was not invariant w.r.t scale. Rousseeuw and Yohai (1984) introduced S-estimator which used the minimization of scale statistics. S-estimator was better than M-estimator as it provided both the location and scale estimates

simultaneously. But, S-estimator was affected by the outlier in multidimensional data and resulted in a significant computational challenge.

Rousseeuw (1985) proposed two estimators; Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD), for approximating the location and scatter of the data and were used for the identification of outliers. In MVE, the location estimate was the center of the ellipsoid and covariance estimate was found by the shape matrix of the ellipse. The minimal volume ellipsoid found in MVE estimator comprised of minimum of m observations, where m is calculated as $[n/2]+1$; where *n* show number of the samples. The MCD estimator calculated the sub-sample of h observations from the data set, whose covariance matrix was found by using minimum optimization problem that means it had the minimum determinant. The mean and the covariance matrix of the h observations were calculated to obtain the mean vector and the covariance estimate respectively. The calculated mean vector and covariance matrix were used to obtain the Mahalanobis distance in order to detect the outliers in the dataset. Numerous algorithms were designed to calculate MCD and MVE estimators. MVE estimate method was a resampling method, as it kept on taking *s* sub-samples of size *m* + 1 from the original data set, *s* was selected to fortify a high probability that at least any one of the sub-sample would be outliers free. For every sub-sample, the mean vector and covariance matrix were calculated, and the volumes of all of the *s* resultant ellipsoids were then estimated, and the one with the least volume, formed the MVE estimate.

Rousseeuw and Leroy (1987) suggested a reweighting step to increase the efficiency of MVE estimator, the step included the recomputation of the mean vector and covariance matrix by using the data points whose squared Mahalanobis distance comparative to the MVE mean vector and covariance matrix was less than the specified Chi-square distribution $\chi^2_{0.5,p}$, here p denoted the degrees of freedom. The MVE estimate required less computational time as compared to the MCD estimator, hence; initially MVE was used for the identification of outlier. Butler (1993) indicated that MCD estimator had improved statistical efficiency as compared to MVE, as MCD was asymptotically normal so it's better to used MCD for the outlier detection. Rousseeuw and Driessen (1999) suggested the FAST-MCD outlier detection method, to approximate the MCD solution. It involved the C-step theorem that states: "if half-sample of data is considered and arranged the complete data set on the basis of Mahalanobis distances, obtained from the mean vector and covariance matrix of the half sample, and then pick a new half-sample from the observations with least distances, if the covariance determinant of the new half-sample would be smaller than or equivalent to the old half-sample covariance determinant, the one with the minimum covariance matrix would be considered for calculating the distance". By repetitively applying this theorem to a dataset, it was possible to converge to at a minimum local optimal MCD solution and calculating the Mahalanobis distance, leads to the identification of outliers.

Hadi (1992) suggested an MVE-based method for the outlier detection based on calculation of the coordinate-wise median vectors for the original dataset and then used the vectors to evaluate the covariance matrix. These location and covariance estimates

were used to calculate the robust Mahalanobis distances for the observations. From this set of distances, the basic subset would be designed from p +1 observations, having the minimum distances. There would be a single basic subset that is composed of observations closed to the centroid of the data that was calculated by the coordinate-wise median robust Mahalanobis distance. The significant decrease in the subsets marked Hadi's method, less computationally difficult and faster to perform.

Billor (2000) proposed a method for outlier detection termed as BACON (Blocked Adaptive Computationally-efficient Outlier Nominator). The method involved the selection of basic subset which was formed from the p+1 observation and had the minimum distances as compared to the component-wise median of the observations. The covariance matrix that was obtained from the median vector and were used to calculate the Mahalanobis distance. These distances were matched to the square root of an appropriate quantile from the χ_p^2 distribution with p degrees of freedom. The use component-wise median marked the BACON technique more robust to outliers at the expense of affine equivariance, since the median estimator was not affine equivariant.

Pena and Prieto (2001) suggested the method for multivariate outlier detection that was based on projecting the data points onto a set of 2p directions and the information was used to maximize and minimize the kurtosis coefficient of the data being projected. Kurtosis is a measure of how peaked the distribution is, which means that if there is high kurtosis, it is heavy tailed or more outliers and low kurtosis means less outliers. The method involved the projecting of data on a vector positioned on the p-

dimensional unit hypersphere and then used the univariate projections of the data and univariate outlier detection to recognize the multivariate outlier in the dataset. For each of 2p directions, there was a backward search algorithm built on the univariate median and MAD, to detect the potential outliers in the data. Based upon the sample mean and covariance of all points not labeled as potential outliers, robust Mahalanobis distances were calculated for all the observations. Those points whose MD> $\chi^2_{0.99,p}$ were considered to be outlier. This process was repeated until the convergence obtained. It aimed to significantly improve computational speed without sacrificing the accuracy of the results. Projection pursuit method could be problematic for higher dimensional datasets since the number of projection vectors created to reach the uniform convergence of the *p*-dimensional unit hyper sphere could grow non-linearly with *p*.

2.9: METHODS FOR OUTLIER DETECTION IN BIVARAITE DATA

Goldberg and Iglewicz (1992) proposed two types of bivariate box plots, termed as "relplot" (a robust elliptic plot) and a "quellplot" (a quarter elliptic plots) for the detection of outliers. The relplot was constructed for the dataset that was assumed to be elliptically symmetrical and ellipses were found by fitting a bivariate Gaussian distribution. Quellplot was used for a non-symmetric data, in which four separate quarter ellipses were found, based on robust estimates of location and scale. The relplot was centered at the mean, whereas quellplot was centered at the center of probability and both the plots showed the location and scale of the data by two intersecting line segments that were either on the regression lines or on the major and minor axes. The interior region of the

plot contained 50% of the data and the observations occurred in the outer region were specified as potential outliers.

Zani et al. (1998) developed a technique for constructing a bivariate boxplot and explained how it might be applied to discover multivariate outliers in the data. The boxplot proposed was based on the method of convex hull and B-spline. The bivariate boxplot for the pair of variables was formed in which the inner region for the plot was the inter-quartile region and determined through the use of convex hull peeling. The method proposed the trimming of data until half of the observations remain and formed the inner region for the boxplot. The method used of B-splines to construct a smooth ellipse that formed the inner region, by fitting a curve to the convex hull of the inner region. The centroid for the boxplot was computed as the arithmetic mean of the observations contained in the inner region. To detect multivariate outliers, construct a bivariate boxplot for every pair of variables. Any observation that was outside the 90% convex hull in any of the plots was removed from the data set. Anthony et al. (1997) discussed about the detection of multiple outliers in bivariate boxplots. For multivariate transformations, the initial subset were found using the contours of the bivariate boxplots and the robust centers were calculated by arithmetic mean of those observations lying within the 50 percent contour. Outliers were detected by calculating the ordered Mahalanobis distances, using the robust centers and covariance matrix.

Rousseeuw et al. (1999) introduced the bagplot to detect the outliers in 2 dimensions. The main idea behind it is half space location depth of a point relative to

bivariate data. The “depth median” was the deepest location and surrounded by a bag containing n/2 observations. The bag was magnified by a factor of 3 and it was named as fence and the observations which lie outside the fence were treated as outlier. The bagplot visualized the location, spread, correlation, skewness and the tails. It could be extended to higher dimension but their algorithms were not yet available. For larger data sets, the computational time for the bagplot increases.

Tongkumchum (2005) introduced a new and simple bivariate boxplot to identify the outliers, which is based on fitting a robust line to the scatter plot of the dataset, and then constructing a box surrounding the fitted line. This two-dimensional box plot contained a pair of trapeziums which were oriented in the direction of a fitted straight line, with the symbol indicating the extreme values. The straight line assigned to the bivariate data set in the boxplot, was named as Tukey’s line. The key components of this two-dimensional box plot were an “inner box” containing 50% of the projection points of observations that were on the fitted line, “a median point” that was inside the inner box, and an “outer box” that separates outliers from the rest of the observations. The two-dimensional box plot visualized the location, spread correlation and skewness of the data.

Sajesh and Srinivasan (2013) discussed about the occurrence of multiple outliers in multidimensional data and the methods used for the identification of hidden outliers in the data set, as growing importance of identification of outliers in a wide variety of practical situations. In their study, they characterized the methods into distance based methods that used to calculate the mean and covariance matrix, using the robust estimates

that were M-estimator, MCD, MVE and BACON and then robust Mahalanobis distances were calculated for each point, using the mean and covariance matrix of the estimator. More research required to find the robust covariance estimates that could be computed efficiently, while keeping the robustness against the outliers. The objective of the non-traditional methods was to found the best projections that could reveal the outliers in an extremely visible situation. Such methodologies could detect the outliers in an extensive range of configurations, as the original location of the outliers was transformed to the more informative projections. However, these methods tend to be very computationally intensive and not suitable for large datasets.

CHAPTER - 3

3. METHODOLOGY

3.1: OVERVIEW

In this study, the algorithm of SSSBB considering bivariate data has been formed. Furthermore; SSSBB is being compared with the Mahalanobis distance for different distributions on the basis of ratio to outlier detected and the area of fence.

3.2: SSSBB-ADIL VERSION

Adil (2011) proposed the 'SSSBB' technique, assuming unimodel and considering both the normal and non-normal data. SSSBB computes information lying on either side of the median ranging from 12.5 percentile to 87.5 percentile of the whole data. In SSSBB, the data is divided into eight parts for the detection of outliers. By calculating the interquartile ranges from left and right side, lower and upper critical values are calculated. First quartile, third quartile and interquartile range from left side of the median gives the lower critical value and from the right side of the median gives the upper critical value. The observations which lie outside the interval of LCV and UCV will be marked as outliers. SSSBB gives the best results as compared to other techniques, when the data is more skewed.

The benefit of applying SSSBB is that the critical values move towards the skewed side of the distribution and covers the actual position of the data. SSSBB

calculates 12.5 and 87.5 percentile and IQR_L and IQR_R that are helpful in determining the fence, whether the distribution of data is right skewed or left skewed.

3.3: EXTENDED SSSBB

Adil (2011) made split sample skewness based boxplot (SSSBB) for univariate normal and skewed data. This thesis extends the SSSBB to bivariate case. The methodology involved in it includes following steps.

1. Calculate the robust regression for the data using robust fit, in order to find out the major axis (the regression line) and minor axis (the line perpendicular to the major axis) through center of data.

2. Find the projection of all data points on the major and minor axis and calculate the distance along major axis (h_1, h_2 … h_n) and minor axis (v_1 v_2 …v_n) for each data point.

3. Apply SSSBB separately on horizontal points that are h_1, h2... hn and on vertical points v1, v2,…,vn to calculate the critical values.

3.3.1: ROBUST REGRESSION

An estimator or statistical procedure is robust, if it provides useful information even if some of the assumptions used to justify the estimation method are not applicable. Robust methods attempt to dampen the influence of outlying cases in order to provide a better fit to the majority of the data. The Robust regression mechanism in MATLAB works by assigning a weight to each data point in the data set. Weighting of the data point is done

iteratively using a method called iteratively reweighted least squares (IRLS) with a bisquare weighting function.

In the first iteration of the IRLS, each point is assigned equal weight and model coefficients are estimated using ordinary least squares. At consequent iterations, weights are recomputed, so that points that are beyond from the model predictions in the previous iteration are given lower weight. Model coefficients are then recomputed using weighted least squares. The process continues until the values of the coefficient estimates converge within indicated tolerance.

3.3.2: CALCULATING HORIZONTAL AND VERTICAL DISTANCES USING ROBUST REGRESSION

Suppose the regression model,

$$Y = \alpha + \beta X + \in$$

and the data points (x_1,y_1), (x_2,y_2),…(x_n,y_n) scattered in xy axis.

The following figure 3.1 illustrates the projection of a single data point (x_1,y_1) in the xy-axis.

Figure 3.1 Projection of a data point along the regression line

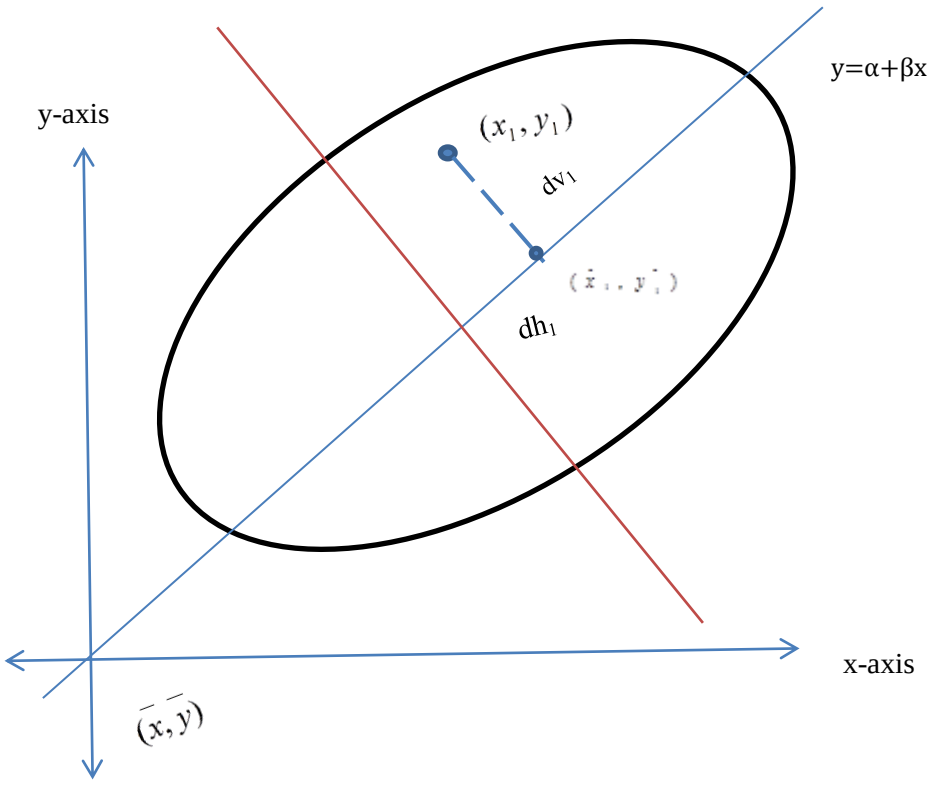


Figure 3.1 shows the projections of the point (x_1, y_1) on the regression line. The major axis is the horizontal line and the minor axis is the vertical line found using the robust fit. The projection of the data point (x_1,y_1) is found on the regression line that is the point $(\dot{x}_1, \dot{y}_1)$ in order to find the horizontal and the vertical distance for the point.

The vertical distance for a point (x_1, y_1) can be calculated using the distance formula as:

$$d_v = \sqrt{(x_1 - \dot{x}_1)^2 + (y_1 - \dot{y}_1)^2}$$

the horizontal distance is calculated using the distance formula as:

$$d_h = \sqrt{(\dot{x}_1 - \bar{x})^2 + (\dot{y}_1 - \bar{y})^2}$$

After calculating the horizontal and vertical distances, critical values for both the distances will be evaluated in order to detect the observations which lie outside the interval.

3.3.3: CRITICAL VALUES FOR HORIZONTAL DISTANCE

To find the critical values for the horizontal distance, SSSBB involves the following steps:

Divide the data into two parts from the median; the lower part and the upper part.

1. By calculating 12.5 percentile, 37.5 percentile, 62.5 percentile and 87.5 percentile the data is divided into eight equal parts.

2. Calculate interquartile range from the upper side and the lower side.

3. For calculating lower critical value, multiply 1.5 into lower interquartile range and subtract it from first quartile of the lower side and for upper critical value multiply 1.5 into upper interquartile range and add into third quartile of the upper side.

Mathematically it is written as:

Q_{1L} = 12.5th percentile,

Q_{3U} = 87.5th percentile,

$IQR_L = Q_{3L} - Q_{1L}$=37.5th percentile - 12.5th percentile,

$IQR_U = Q_{3R} - Q_{1R}$ = 87.5th percentile - 62.5th percentile

As left and right interquartile ranges are calculated separately, so skewness will be handled automatically. Lower and upper boundaries are defined as

$$[LCV_h \quad UCV_h] = [Q_{1L} - 1.5 * IQR_L \qquad Q_{3U} + 1.5 * IQR_U]$$

Where L shows the lower critical value and U represents the upper critical value for the skewed distribution. An observation which lies outside of this limit is considered as outlier.

3.3.4: CRITICAL VALUES FOR VERTICAL DISTANCE:

For finding the upper and lower critical value for vertical distance, SSSBB will be calculated separately. It involves the same steps as for the horizontal distance. The critical values for the vertical distance will be calculated and used for the detection of outliers.

Divide the data into two parts from the median that is the lower part and the upper part.

1. By calculating 12.5 percentile, 37.5 percentile, 62.5 percentile and 87.5 percentile the data is divided into eight equal parts.

2. Calculate interquartile range from the upper side and the lower side.

3. For calculating lower critical value multiply 1.5 into lower interquartile range and subtract it from first quartile of the lower side and for upper critical value multiply 1.5 into upper interquartile range and add into third quartile of the upper side.

For vertical distance the interval will be:

$$[LCV_v \quad UCVv] = [Q_{1L} - 1.5 * IQR_L \qquad Q_{3U} + 1.5 * IQR_U]$$

The vertical distance values which lie outside this interval will be marked as outliers.

3.3.5: AREA OF FENCE

After applying SSSBB to the horizontal and vertical distance, the area of the fence will be calculated to identify the points which lie outside the specified area. As, it makes a rectangular shape, so the area is calculated as:

$$AreaSSSBB = (UCV_h - LCV_h)(UCV_v - LCV_v)$$

3.3.6: OUTLIER DETECTION

The observations which lie outside the region will be marked as outliers, as there distances are greater than the critical values, as calculated in horizontal and vertical distances.

3.4: MAHALANOBIS DISTANCE

A point is said to be outlier if it lies sufficiently far away from the remaining data set. To calculate the distance of the point the famous technique is robust distance.

For the data set X it is defined as

$$RD_i = \sqrt{(x_i - \mu_{MCD})^t \Sigma_{MCD}^{-1} (x_i - \mu_{MCD})}$$

for each point x_i in the data. μ_{MCD}, it is the MCD estimate of the location and Σ is the MCD covariance estimate.

3.4.1: MINIMUM COVARIANCE DETERMINANT (MCD)

Minimum covariance estimate (MCD), is the robust estimate of location and scatter of multivariate data. It can be computed by the fast algorithm of Rousseeuw and Van Driessen (1999). The classical Mahalanobis distance has few short comings. It is affected by the masking effect that means it cannot detect the regular outliers. But the robust distance can resist the outliers for the reliable data analysis. Robust distance can be computed by the fast minimum covariance determinant. MCD estimator is applicable for elliptically symmetric unimodal distributions. The FAST-MCD algorithm is based on the concept of C-step.

C-STEP:

Consider a data set X = {$x_1 \ldots x_n$} and let $H_1 \subset \{1,\ldots, n\}$ be a h-subset, that is $|H_1| = h$. μ and $\sum_1$ are the empirical mean and covariance matrix of the data in H_1. If det $(\Sigma_1) \neq 0$, the relative distance is defined as

$$d_1(i) = \sqrt{(x_i - \mu_1)^t \Sigma_1^{-1} (x_1 - \mu_1)} \qquad \text{i=1. . . n}$$

Then take the second subset H_2 such that $\{d_1(i); i \in H_1\} := \{(d_1)_{1:n}, \ldots, (d_1)_{h:n}\}$ where $(d_1)_{1:n} \leq (d_1)_{2:n} \leq \cdots \leq (d_1)_{n:n}$ are the ordered distances and compute μ_2 and $\sum_2$ based on H_2. Then,

$$\det(\sum_2) \leq \det(\sum_1)$$

Equality will hold only if the mean and the covariance matrix of both the subsets are equal. If det $(\sum_1) > 0$, the C-step yields a new h-subset which has lower covariance determinant, C stands for the 'concentration'. $\sum_2$ is more concentrated (has a lower determinant) than $\sum_1$. The condition det $(\sum_1) \neq 0$ in the C step theorem is not a real restriction because if det$(\sum_1) = 0$ the minimal objective value has already been reached.

C-steps can be iterated until det$(\sum_{new}) = 0$ or det$(\sum_{new})$ = det$(\sum_{old})$.

There is no guarantee that the final value det$(\sum_{new})$ of the iteration process obtained is the global minimum value of the MCD objective function.

3.4.2: AREA OF FENCE

After applying MCD algorithm, the Mahalanobis distances will be calculated and plotted. It will be in elliptical form and the area will be found easily using area of ellipse formula. Spruyt (2014) described that eigenvalues of the covariance matrix are computed in order to find the major and minor axis of the ellipse.

The length of the major and minor axis is calculated as follows:

$$majoraxis = 2\sqrt{7.378\lambda_1} \quad , \quad \min oraxis = 2\sqrt{7.378\lambda_2}$$

Here, λ_1 and λ_2 represents the eigenvalues of the covariance matrix.

Then, the area of the ellipse is found as:

$$areamd = \pi * majoraxis * \min oraxis$$

3.4.3: OUTLIER DETECTION

By plotting the DD-plot, outliers will be detected using the cutoff value $\chi^2_{p,0.975}$ based on the asymptotic distribution of the robust distances. The values which lie outside the cutoff value will be considered as outliers.

MCD procedure is very fast for small sample sizes *n*, but when *n* grows the computational time increases because of the *n* distances that need to be calculated in each C-step.

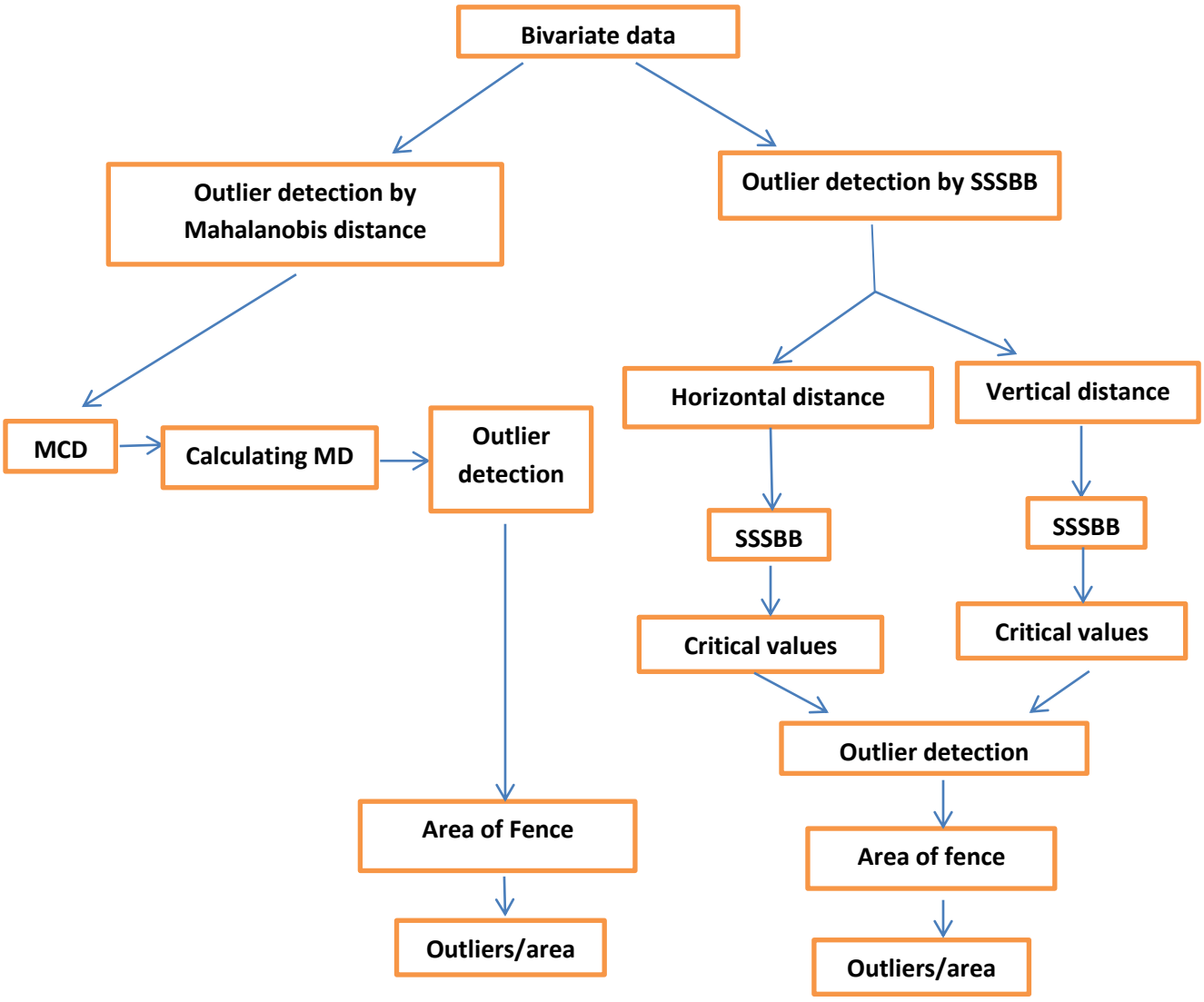
3.5: MONTE CARLO DESIGN

Considering bivariate data, the robust Mahalanobis distance and SSSBB techniques will be used and compared. Firstly outliers will be detected by applying the fast MCD algorithm and robust Mahalanobis distance will be calculated for all the data points. Outliers will be detected using chi-square value at 97.5 percentile and area of the ellipse has found. The observations which lie outside the ratio of outlier detected and area of ellipse are marked as outliers.

SSSBB for bivariate data is introduced to detect the outliers by calculating the horizontal and vertical distances and the SSSBB technique is applied on both the

distances separately to evaluate the critical values in order to detect the possible outliers in the data set. The figure 3.2 shows the Monte Carlo design of the study as follows:

Figure 3.2 Monte Carlo design of the study



3.6 DATA GENERATING PROCESS

We want to analyze the performance of extended SSSBB for a variety of bivariate distribution. The bivariate data can easily be generated as two separate columns from the specific distribution. However, we will take the real data in which some correlation occurs to check the performance of the SSSBB and Mahalanobis distance. Using data generating process (DGP), as pair of variables are treated separately; the correlation pattern in case of skewed distribution is handled by using cholesky decomposition and the process is as follows:

Data using the χ^2 distribution is generated in the following way:

1. Generate X=$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ such that Xi~χ^2(X) and Cov(x_1 x_2) =0.

2. Assume variance covariance matrix , $\sum = \begin{bmatrix} \sigma_{x^2} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{y^2} \end{bmatrix}$

3. Find L such that LL'= $\sum$ by using Cholesky decomposition.

4. Find Y=LX.

5. Repeat 1-4 ‘n’ times to generate nx2 matrix of random variables with cov σ_{xy}.

The change in degree of freedom of χ^2 distribution can change the skewness, so that performance can be evaluated for various measure of skewness. Data for other asymmetric distributions such as beta, gamma is also generated using the DGP steps 1-5, for comparing the results of the techniques: Mahalanobis distance and SSSBB. For the

case of symmetric distribution, DGP is used to generate the data and the validity of SSSBB is checked for t-distribution and its results are compared with Mahalanobis distance.

i) Student t-distribution:

The t-distribution is theoretical probability distribution that is symmetrical, bell shape (close to standard normal distribution) and it is a fat tail distribution.

The probability density function (pdf) of the t-distribution is given as:

$$f_r(t) = \frac{\Gamma[\frac{1}{2}(r+1)]}{\sqrt{r\pi}\Gamma(\frac{1}{2}r)(1+\frac{t^2}{r})^{(r+1)/2}} \qquad , \qquad r = n-1$$

It has a parameter, r, called degree of freedom, denoted as, df, which can be any real number greater than zero. The change in df of the t-distribution, changes the shape of the distribution therefore, with smaller value of df, the graph is flatter and more area is in the tails of the distribution and with the larger df, the area in the tails decreases and the area near the center increases. With the increase in the degree of freedom of t-distribution, it approaches to standard normal distribution. The t-distribution is used in hypothesis testing, to figure out whether to reject or accept the null hypothesis. Therefore, it is important to apply SSSBB on t-distribution and compare its results with Mahalanobis distance.

To check the performance of SSSBB for skewed distribution and to compare its results with Mahalanobis distance, the distributions considered are: chi-square distribution, gamma distribution and beta distribution.

ii) CHI-SQUARE DISTRIBUTION:

The chi-square distribution is the distribution of the sum of squared standard normal deviates that is random sample from the standard normal distribution and its degree of freedom is equal the sum of number of standard normal deviates.

The probability distribution function of chi-square distribution is:

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, x \in (0, \infty)$$

Chi-square distribution is not symmetrical but rather has a positive skewed and its shape, center and spread changes with the change in the degree of freedom. As the number of degree of freedom changes, the skewness of the distribution changes.

It's important to test SSSBB on chi-square distribution as many test statistics like tests of deviations of differences between theoretically expected and observed frequencies (one-way table) and contingency table, are approximately distributed as chi-square. The chi-square distribution is used in many cases for the critical regions for hypothesis tests and in determining confidence intervals.

Two common examples are the chi-square test for independence in an RxC contingency table and the chi-square test to determine if the standard deviation of a population is equal to a pre-specified value.

iii) GAMMA DISTRIBUTION:

Gamma distribution is a right-skewed probability distribution and its pdf is given as:

$$f(x) = \frac{(\frac{x-\mu}{\beta})^{\gamma-1} \exp(-\frac{x-\mu}{\beta})}{\beta\Gamma(\gamma)} \quad , \qquad x \geq \mu; \gamma, \beta > 0$$

The two parameters α and β, defines the shape of the graph.

The parameter α is the shape parameter and β is the rate parameter, they both effect the shape of the graph. The range of the distribution is from (0,∞). The gamma distribution can be used in a range of disciplines including queuing models, climatology, and financial services. The gamma distribution is also used to model errors in a multi-level Poisson regression model because the combination of a Poisson distribution and a gamma distribution is a negative binomial distribution.

iv) BETA DISTRIBUTION:

The beta distribution with left parameter $a \in (0,\infty)$ and right parameter $b \in (0,\infty)$ is the continuous distribution on (0,1) with probability density function given as:

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, 0 < x < 1$$

Here B(a,b) is a beta function and a and b are the positive shape parameters of the distribution.

The beta distribution is used for many applications, including Bayesian hypothesis testing, the Rule of Succession (a famous example being Pierre-Simon Laplace's treatment of the sunrise problem), and Task duration modeling. The beta distribution is especially suited to project/planning control systems like PERT and CPM because the function is constrained by an interval with a minimum (0) and maximum (1) value.

3.5: EMPIRICAL ANALYSIS

The empirical analysis shall be based on the results obtained through the iterative procedure of Mahalanobis distance and SSSBB for the detection of outliers. SSSBB for bivariate data will also be applied on some real world data to have the empirical evidence.

3.5.1: VARIABLES

1) The empirical study will be conducted for Pakistan's Stock Exchange data measured on monthly frequency and both methods are applied for the identification of outliers.

The data will be taken from business recorder site for the time period of July, 2009 – May, 2017. As we consider bivariate data so, closing point and turn over point will be considered for the companies, it will include the following companies: Pakistan

Petroleum Ltd, United Bank Limited, Lucky Cement Ltd, Engro Corporation Limited and Pakistan Oilfields Ltd.

2) The performance of SSSBB is checked on the bivariate data, for which measures of interest rate data that are money market rates and treasury bill rates are considered.

Money market rates:

A money market account is an interest-bearing account that typically pays a higher interest rate than a savings account, and which provides the account holder with limited check-writing ability.

Treasury bill rates:

These are government bonds or debt securities with maturity of less than a year. T- Bills are issued to meet short-term mismatches in receipts and expenditure. Bonds of longer maturity are called dated securities.

3.6: BASIS OF COMPARISON

SSSBB technique for bivariate case will be compared with existing outlier detection method that is robust Mahalanobis distance on the basis of ratio of outliers detected and the area of fence. If there is whole area A where all the data points are scattered, there is a region R inside it where outliers don't exist. Using the critical values the observations which lie outside the area R, will be considered as the outlier. The results of SSSBB and

Mahalanobis distance are based on the observed values of the quartiles (octiles). In usual method, as the area of the region increases, less outlier will be detected. But in a better method, all possible outliers will be detected while keeping the area same.

CHAPTER: 4

4. RESULTS AND DISCUSSIONS

As described in chapter 3, Mahalanobis distance is calculated using minimum covariance determinant (MCD) algorithm, which uses the concept of C-step for finding the subset having minimum determinant value, the distance is calculated using the mean and the minimum covariance matrix. The observations whose value exceeds the critical value of $\chi^2_{0.975}$ are labeled as outliers and the ratio of the outlier detected and area of fence is found for the technique. SSSBB, for bivariate case, algorithm works by calculating the horizontal and vertical distances that is the major axis and minor axis and applying SSSBB on both the distances separately. SSSBB works by dividing data into two parts from the median and finding lower critical value by q_1 and IQR_L and for upper critical value q_3 and IQR_R are calculated. For the outlier detection, the observations which lie outside the interval of critical values are labeled as outliers, the ratio of outlier detected and the area of the fence is calculated. The technique, whose ratio is greater than the other one, is considered to be a good method, as it will detect possible outliers in the specified area.

After developing an algorithm of SSSBB bivariate case, for the identification of outliers, it is necessary to check its performance on variety of distributions and data types. We used MATLAB to compute thousands of simulations for diverse distributions and different parameter values. Both symmetric and skewed distributions are taken to

compare the results of the techniques, SSSBB and Mahalanobis distance. The distributions considered for the estimations are: t-distribution, chi-square distribution, gamma distribution and beta distribution and results of outlier detected, area of fence and ratio, are obtained for both the techniques. The data is generated randomly from Data Generating Process (DGP) described in chapter 3. For each DGP, thee simulations are performed for three different sample sizes: small, medium and large that is 40, 80 and 200, respectively, and for each sample size,10,000 simulations were performed and complied to compute the ratio of outlier detected and the area by both the techniques considered under the comparison. Results are being complied to check whether SSSBB is a good estimation technique or a bad one.

4.1: THE STUDENT t-DISTRIBUTION

If df of the t-distribution is small, we have fat tail t-distribution and if df is large, student's t-distribution converges to normal distribution. Therefore, analyzing performance for t-distribution with various df can indicate the effect of fat tails on the performance of procedure under consideration. A sample from student t-distribution of three different sizes: 40, 80 and 200 are used with different degree of freedom, to check the performance of SSSBB and Mahalanobis distance. As with the change in degree of freedom in t-distribution the size of the tail changes, if the estimator maintains its performance with the change in the tail then it is good estimator otherwise it is not a preferable estimator.

Table 4.1 Performance of SSSBB and Mahalanobis distance for student t-distribution

SAMPLE SIZE(SS)		SSSBB			MAHALANOBIS DISTANCE		
		OD	**AREA**	**RATIO**	**OD**	**AREA**	**RATIO**
SS=40	df(1)	8.30	186.36	0.07	11.54	635.49	0.02
	df(2)	5.21	50.26	0.12	7.04	316.15	0.02
	df(5)	2.90	26.15	0.13	4.28	204.40	0.02
	df(10)	2.24	21.69	0.12	3.42	176.15	0.02
	df(15)	2.06	20.38	0.12	3.20	167.43	0.02
SS=80	df(1)	17.82	165.08	0.13	22.91	623.36	0.04
	df(2)	10.80	49.98	0.23	13.59	317.54	0.04
	df(5)	5.63	26.50	0.23	7.73	206.03	0.04
	df(10)	4.27	21.85	0.21	5.99	176.47	0.03
	df(15)	3.83	20.69	0.20	5.48	167.68	0.03
SS=200	df(1)	45.60	156.90	0.33	57.03	619.10	0.09
	df(2)	27.46	49.90	0.57	33.45	318.70	0.11
	df(5)	13.86	26.59	0.54	18.68	205.83	0.09
	df(10)	10.10	22.07	0.47	14.19	176.14	0.08
	df(15)	9.04	20.81	0.45	12.76	167.14	0.08

OD: Outlier Detected, AREA: Area of fence, RATIO= OD/Area of fence

Figure 4.1 Performance of SSSBB and Mahalanobis distance for student t-distribution

Figure 4.1(a)

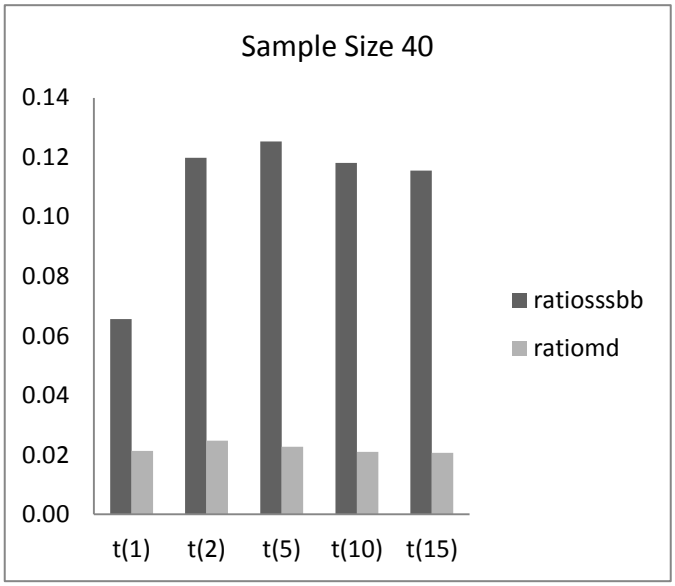


Figure 4.2 (b)

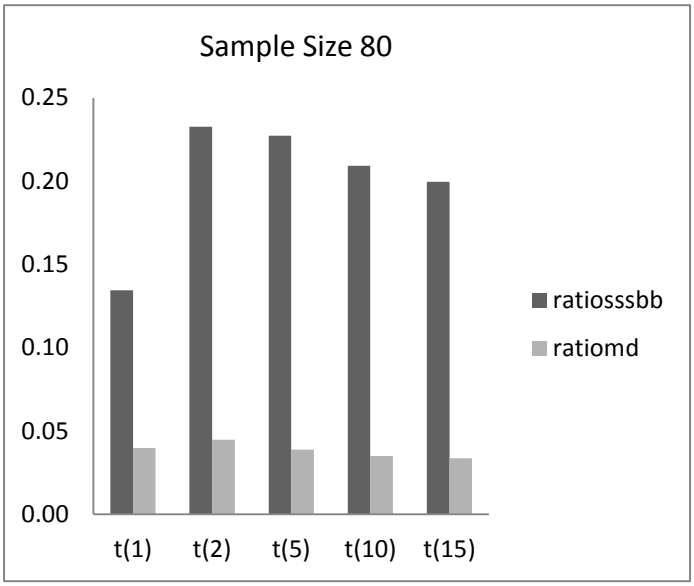


Figure 4.1(c)

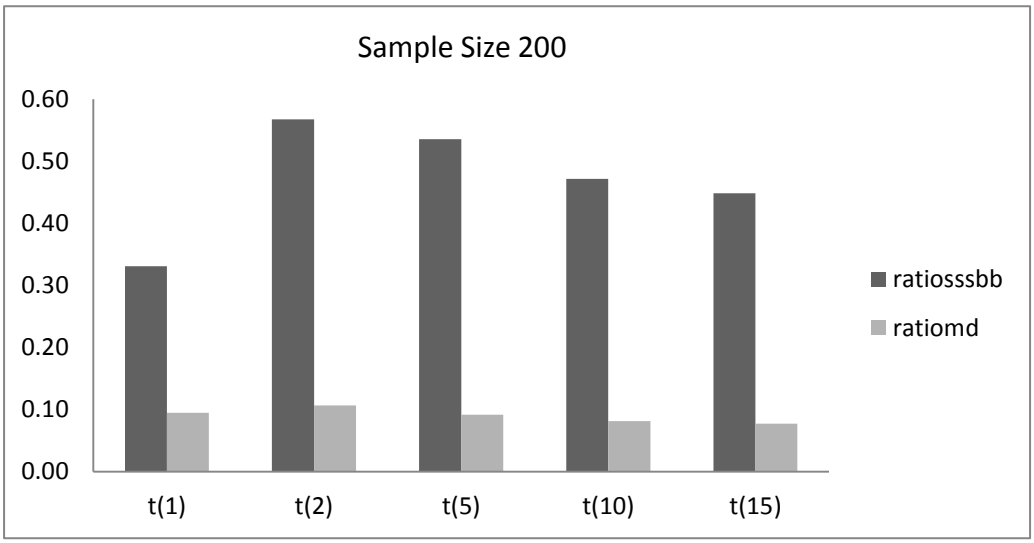


Table (4.1) shows the results of t-distribution in which the averages of the simulations results of outlier detected; area and ratio of SSSBB and Mahalanobis distance are calculated with degree of freedom 1, 2,5,10 and 15. Row 1 of the top panel of the table indicates that for sample size 40 and df=1, the ratio of outlier detected and the area for SSSBB is 0.07 and for Mahalanobis distance the ratio is 0.02, that is less than SSSBB. Row 2 of the top panel shows the ratio of SSSBB for df=2 is 0.12 that is far more than the ratio of Mahalanobis distance which is 0.02. Therefore as the df for t-distribution increases, it approaches to normality. When the df value is small, the t-distribution covers more area in the tails; as a result, more extreme values will come under the t-distribution.

The results are computed for the different values of degree of freedom for t-distribution. The performance of SSSBB is better than the Mahalanobis distance as with the change in the value of df, the ratio of SSSBB is greater as it detects the possible outliers within the specified area while Mahalanobis distance detects more outliers, covering more area. Figure 4.1(a) shows that for small sample size 40 and at different degree of freedoms the performance of SSSBB is far better than Mahalanobis distance in t-distribution, the greater ratio of SSSBB shows that more possible outliers are detected with less area while Mahalanobis distance detect less outliers covering more area as compared to SSSBB. The df is on x-axis while ratios are on y-axis. The greater ratios of SSSBB proved it to be a better method as compared to Mahalanobis distance.

Figure 4.1 (b) shows the graph for t-distribution with medium sample size 80, it is cleared from the graph that SSSBB performs well as compared to Mahalanobis distance. Figure 4.1(c) shows graph for a large sample size of 200, with different degree of freedom, in which the ratio of SSSBB is more than Mahalanobis distance. Therefore, simulation results and their graphs shows that SSSBBB performance is better than the Mahalanobis distance in approximately normal distribution.

4.2: CHI-SQUARE DISTRIBUTION

The performance of SSSBB is checked for skewed distribution: chi-square distribution and compared the results with Mahalanobis distance on the basis of ratio of outlier detected and the area of fence. As the degree of freedom of chi-square distribution changes, the skewness of the graph changes. If df of the χ^2 distribution is small, the skewness is high but as the df increases, skewness decreases and approaches to normal distribution. Therefore, to check the effect of skewness with outlier detection methods considered in the study, we will take various values of df for χ^2 distribution. If SSSBB, maintains its performance with the change in the degree of freedom then it is a good method as compare to Mahalanobis distance.

Table 4.2 Performance of SSSBB and Mahalanobis distance for chi-square distribution

SAMPLE SIZE(SS)		**SSSBB**			**MAHALANOBIS DISTANCE**		
		OD	AREA	RATIO	OD	AREA	RATIO
SS=40	df(2)	2.51	60.90	0.21	8.15	308.08	0.12
	df(5)	2.11	169.77	0.06	5.08	1184.64	0.02
	df(10)	1.92	351.05	0.02	3.95	2690.36	0.01
	df(15)	1.88	530.83	0.02	3.53	4197.45	0.00
	df(20)	1.88	714.13	0.01	3.34	5723.20	0.00
SS=80	df(2)	5.09	62.51	0.09	15.24	312.04	0.05
	df(5)	3.93	172.21	0.02	9.17	1188.20	0.01
	df(10)	3.59	355.95	0.01	6.99	2699.26	0.00
	df(15)	3.45	539.15	0.01	6.16	4208.74	0.00
	df(20)	3.41	722.53	0.01	5.76	5723.34	0.00
SS=200	df(2)	12.69	63.42	0.21	36.05	311.49	0.12
	df(5)	9.42	173.43	0.06	21.57	1188.19	0.02
	df(10)	8.40	358.56	0.02	16.28	2694.93	0.01
	df(15)	7.97	544.79	0.02	14.33	4203.76	0.00
	df(20)	7.79	729.80	0.01	13.32	5701.45	0.00

Figure 4.2 Performance of SSSBB and Mahalanobis distance for chi-square distribution

Figure 4.2(a)

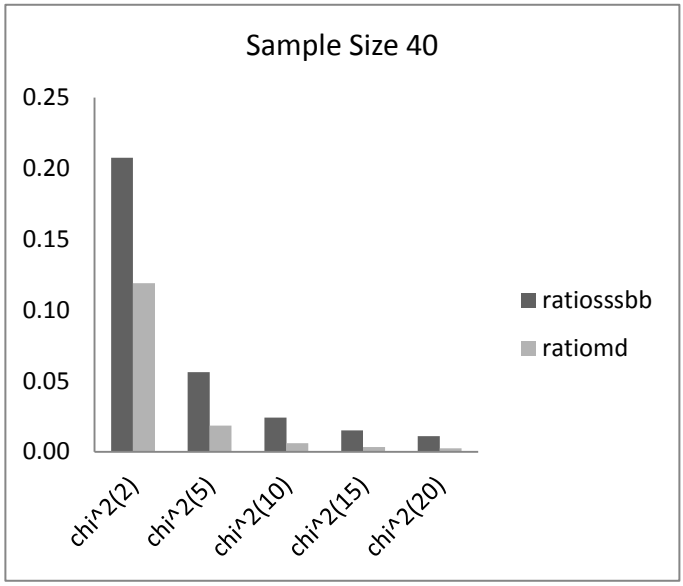


Figure 4.2(b)

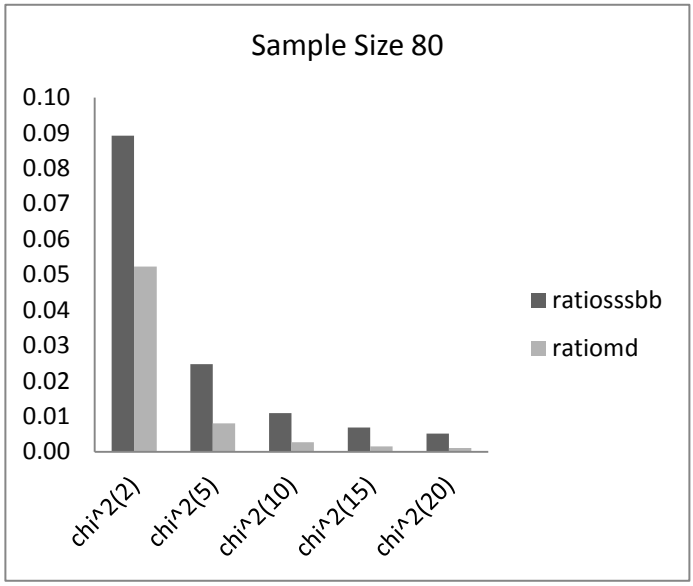
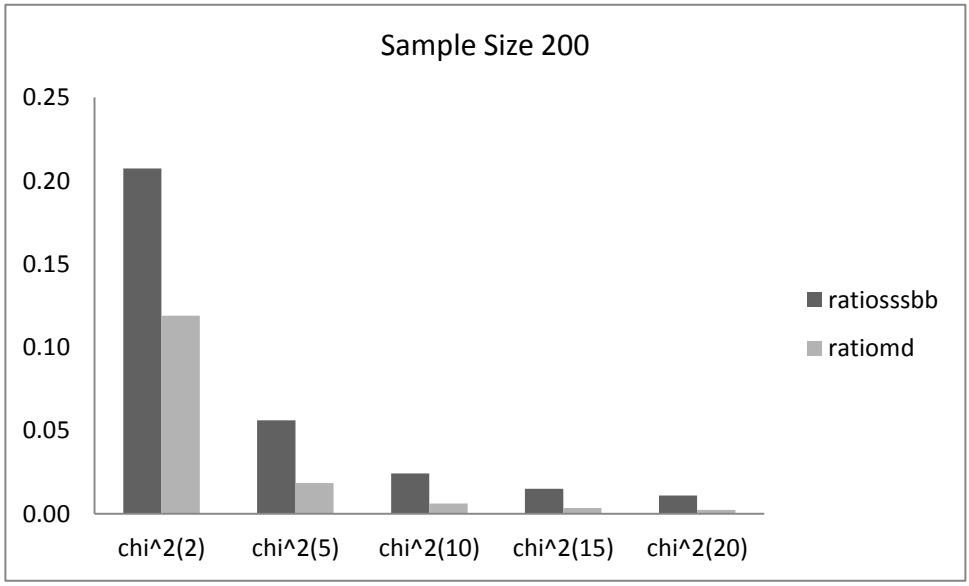


Figure 4.2(c)



Table(4.2) shows the simulation results of chi-square distribution with sample sizes:40,80 and 200 and degree of freedom(df): 2,5,10,15,20. For these values of df, SSSBB and Mahalanobis distance is evaluated, outlier detected and the area covered by the techniques. Row 1 of the top panel of the table shows that with df=2 and sample size 40, SSSBB ratio is 0.21 while Mahalanobis distance ratio is 0.12 therefore SSSBB ratio is greater. Row 2 indicates that SSSBB, applied on chi-square distribution for df=5 detects the possible outliers under the specified area, the ratio is 0.06 while Mahalanobis distance detects more outliers covering large area and its ratio is 0.02, therefore greater value of SSSBB shows that it's a good method.

Results shows that SSSBB performs well as compared to Mahalanobis distance, as the ratio of SSSBB is greater than the Mahalanobis distance, which means SSSBB covering small area and detect the possible outliers in the sample.

Figure 4.2(a) shows the graph of ratio SSSBB and ratio Mahalanobis distance using chi-square distribution for sample size 40, as the degree of freedom changes the skewness changes, so that the performance of the technique is measured for different degree of freedom. At df(2), SSSBB ratio is more than the Mahalanobis distance which means SSSBB detects all the possible outliers covering small area, while Mahalanobis distance have greater area of fence and detecting less outliers.

For df= 5, 10, 15 and 20, ratio of outlier detected to the area of fence for SSSBB is greater than the Mahalanobis distance. Figure 4.2 (b) is for the medium sample size that is 80; it shows that, SSSBB is a good method for the outlier detection, it detect the possible outliers while Mahalanobis distance, covering more area, is unable to detect the possible outliers.

Figure 4.2 (c) shows the graph of chi-square distribution for sample size 200, with different degree of freedoms to check the performance of SSSBB. The greater ratio of SSSBB shows that it is a good method as compare to Mahalanobis distance.

4.3: GAMMA DISTRIBUTION

Gamma distribution is a right-skewed probability distribution with two parameters α and β, which defines the shape of the graph. The parameter α is the shape parameter and β is the rate parameter, they both effect the shape of the graph. The performance of SSSBB is checked for gamma distribution and 10000 simulations are run to have the results. The ratio of outlier detected and area of fence for SSSBB and Mahalanobis distance is calculated to find the preferable technique.

Table 4.3 Performance of SSSBB and Mahalanobis distance for gamma distribution

SAMPLE SIZE(SS)		SSSBB			MAHALANOBIS DISTANCE		
		OD	AREA	RATIO	OD	AREA	RATIO
SS=40	(α,β)=(2,1)	2.18	33.15	0.08	5.57	221	0.03
	(α,β)=(2,1.5)	2.15	74.66	0.03	5.58	495.49	0.01
	(α,β)=(3,1)	2.04	51.38	0.05	4.67	370.78	0.01
	(α,β)=(3,1.5)	2.03	115.57	0.02	4.65	836.17	0.01
	(α,β)=(4,1)	1.95	69.6	0.03	4.21	520.76	0.01
	(α,β)=(4,1.5)	1.97	157.01	0.01	4.19	1178.71	0
	(α,β)=(5,1)	1.93	87.73	0.03	3.95	671.4	0.01
	(α,β)=(5,1.5)	1.94	197.35	0.01	3.91	1516.35	0
SS=80	(α,β)=(2,1)	4.16	33.82	0.13	10.26	222.32	0.05
	(α,β)=(2,1.5)	4.16	76.07	0.06	10.25	500.37	0.02
	(α,β)=(3,1)	3.87	51.94	0.08	8.47	371.87	0.02
	(α,β)=(3,1.5)	3.84	117.05	0.04	8.47	836.85	0.01
	(α,β)=(4,1)	3.7	70.52	0.06	7.52	524.97	0.01
	(α,β)=(4,1.5)	3.72	158.06	0.03	7.45	1178.8	0.01
	(α,β)=(5,1)	3.6	89	0.04	6.96	675.68	0.01
	(α,β)=(5,1.5)	3.61	200.22	0.02	6.98	1522.94	0
SS=200	(α,β)=(2,1)	10.06	34.17	0.3	24.15	222.25	0.11
	(α,β)=(2,1.5)	10	76.79	0.13	24.18	499.06	0.05
	(α,β)=(3,1)	9.09	52.63	0.18	19.9	372.25	0.05
	(α,β)=(3,1.5)	9.11	118.36	0.08	19.89	837.71	0.02
	(α,β)=(4,1)	8.59	71.14	0.12	17.63	523.15	0.03
	(α,β)=(4,1.5)	8.65	160.15	0.06	17.6	1178.06	0.02
	(α,β)=(5,1)	8.33	89.62	0.1	16.21	674.08	0.02
	(α,β)=(5,1.5)	8.34	201.81	0.04	16.26	1517.46	0.01

Figure 4.3 Performance of SSSBB and Mahalanobis distance for Gamma distribution

Figure 4.3(a)

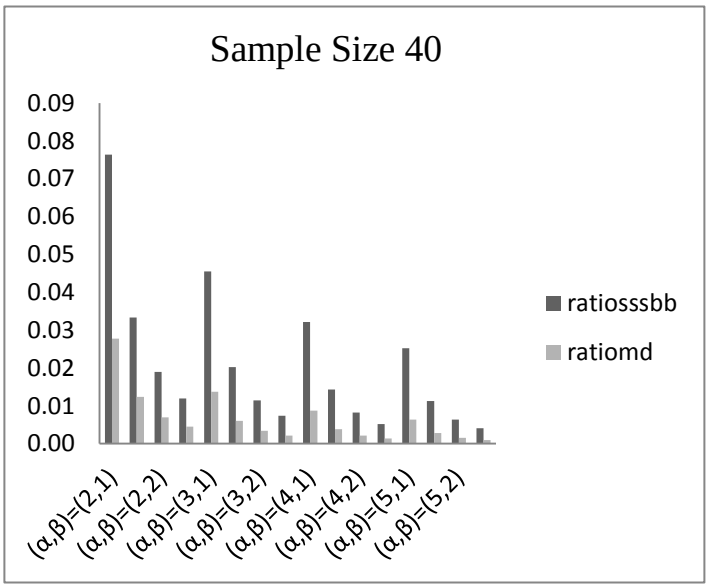


Figure 4.3(b)

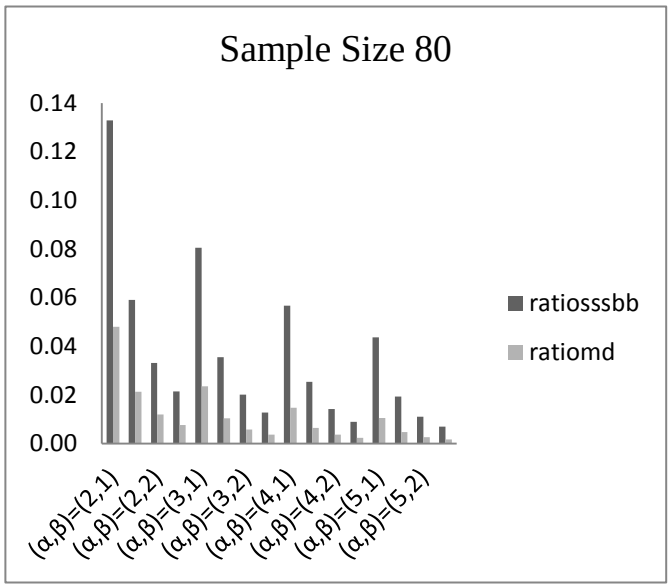


Figure 4.3(c)

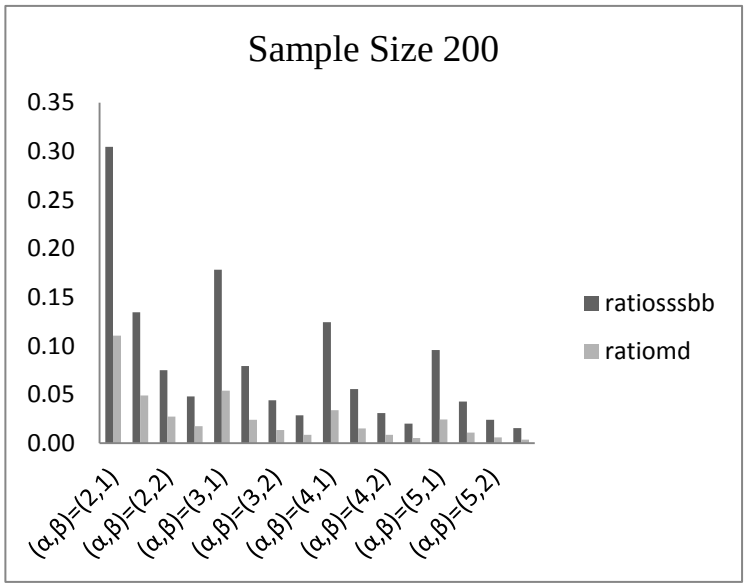


Table (4.3) shows the simulation results of SSSBB and Mahalanobis distance performed on gamma distribution with different α and β values, with the change in the values of these parameters the shape of the graph changes therefore it is important to check the performance of both the techniques for the different parameter values, for α are 2, 3, 4 and 5 and for β are 1, 1.5, 2 and 2.5. As shown in table row 1 of the top panel, with α=2 and β=1 and sample size 40, the average outliers detected by SSSBB are 2.18 and the area of fence is 33.15, so the ratio calculated is 0.08 while Mahalanobis distance with average 5.57 outliers detected and area covered 221, the ratio is 0.03, which is less than the SSSBB ratio. Row 2 shows that gamma distribution with α=2 and β=1.5, the ratio of SSSBB is 0.03 while Mahalanobis distance has the ratio 0.01; therefore SSSBB has the greater ratio as compared to Mahalanobis distance.

Taking different parameter values for the gamma distribution, SSSBB and Mahalanobis distance are applied. As shown in the table the results for SSSBB are far better than the Mahalanobis distance, which means that SSSBB detects the possible outliers within the specified area therefore it has greater ratio as compared to Mahalanobis distance.

Figure 4.3(a) shows the graph of gamma distribution for small sample size 40, the ratio of SSSBB is greater than the Mahalanobis distance; therefore, it is a good method for outlier detection. For sample size 80, figure 4.3(b) shows that SSSBB detect the possible outliers with less area of fence, as its ratio is greater than the Mahalanobis

distance while Mahalanobis distance has greater area of fence and detect more outliers. Again for large sample size 200, both the methods are applied and simulations are run to find the good method. As shown in figure 4.3(c), the ratio of SSSBB is greater than the Mahalanobis distance and it maintains its performance with the change in the values of the shape parameters.

4.4: BETA DISTRIBUTION

Beta distribution is used to check the performance of SSSBB and Mahalanobis distance and on the basis of ratio of outlier detected and area of fence, preferable method is proposed. The method whose ratio is greater than the other one is a good one and has more capability to detect the possible outliers in the data set. In beta distribution, fat tail and skewness both are changed with the change in the values of the parameters α and β, these parameters are responsible for the change in the shape of the graph and are positive shape parameters. SSSBB and Mahalanobis distance are applied on the beta distribution for different sample sizes using different parameters values, for α are 2, 5, 10, 15 and for β are 2, 4, 8, 10 and 10000 simulations are run to have the results for the beta distribution.

Table 4.4 Performance of SSSBB and Mahalanobis distance for beta distribution

SAMPLE SIZE(SS)		SSSBB			MAHALANOBIS DISTANCE		
		OD	AREA	RATIO	OD	AREA	RATIO
SS=40	(α,β)=(2,2)	1.06	0.96	1.27	1.68	8.23	0.22
	(α,β)=(2,4)	1.35	0.60	2.61	2.78	4.91	0.61
	(α,β)=(5,2)	1.46	0.47	3.59	3.24	3.79	0.92
	(α,β)=(5,4)	1.39	0.46	3.48	2.22	3.90	0.61
	(α,β)=(10,2)	1.77	0.19	10.83	4.34	1.42	3.30
	(α,β)=(10,4)	1.59	0.25	7.36	2.87	2.05	1.50
	(α,β)=(15,2)	1.87	0.10	21.42	4.73	0.73	7.03
	(α,β)=(15,4)	1.69	0.15	12.88	3.20	1.22	2.82
SS=80	(α,β)=(2,2)	1.58	0.98	1.77	2.24	8.07	0.28
	(α,β)=(2,4)	2.22	0.60	4.01	4.41	4.89	0.93
	(α,β)=(5,2)	2.51	0.48	5.70	5.37	3.81	1.45
	(α,β)=(5,4)	2.37	0.47	5.47	3.24	3.87	0.86
	(α,β)=(10,2)	3.18	0.19	17.78	7.62	1.44	5.49
	(α,β)=(10,4)	2.78	0.25	11.85	4.71	2.05	2.36
	(α,β)=(15,2)	3.48	0.10	36.80	8.46	0.73	11.99
	(α,β)=(15,4)	3.06	0.15	21.57	5.49	1.22	4.64
SS=200	(α,β)=(2,2)	2.81	0.99	2.98	4.44	7.78	0.57
	(α,β)=(2,4)	4.59	0.61	6.37	9.49	4.85	1.97
	(α,β)=(5,2)	5.28	0.49	11.27	11.83	3.80	3.14
	(α,β)=(5,4)	4.92	0.47	10.78	7.00	3.79	1.86
	(α,β)=(10,2)	7.31	0.20	38.67	17.46	1.44	12.30
	(α,β)=(10,4)	6.14	0.26	24.88	10.62	2.03	5.28
	(α,β)=(15,2)	8.17	0.10	81.58	19.63	0.74	27.06
	(α,β)=(15,4)	6.75	0.15	45.27	12.54	1.21	10.47

Figure 4.4 Performance of SSSBB and Mahalanobis distance for beta distribution with different sample sizes

Figure 4.4(a)

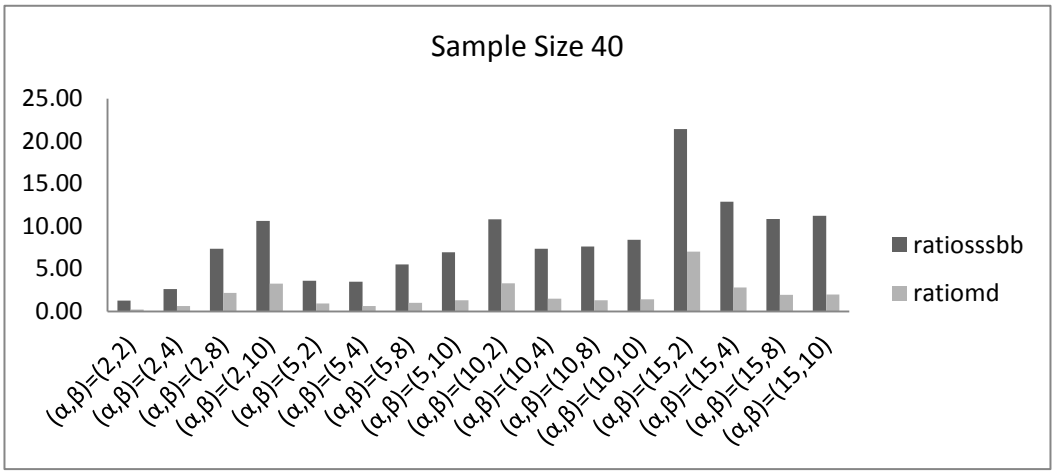


Figure 4.4(b)

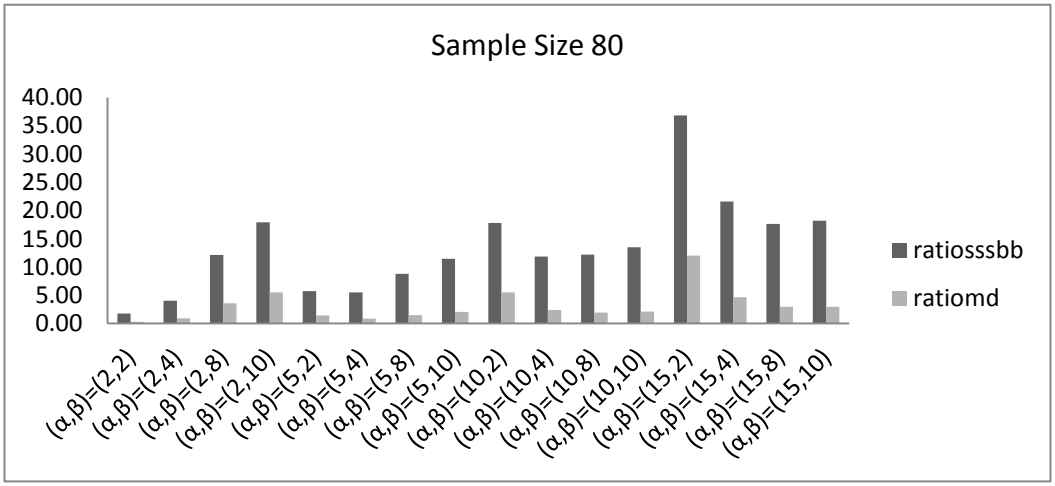


Figure 4.4(c)

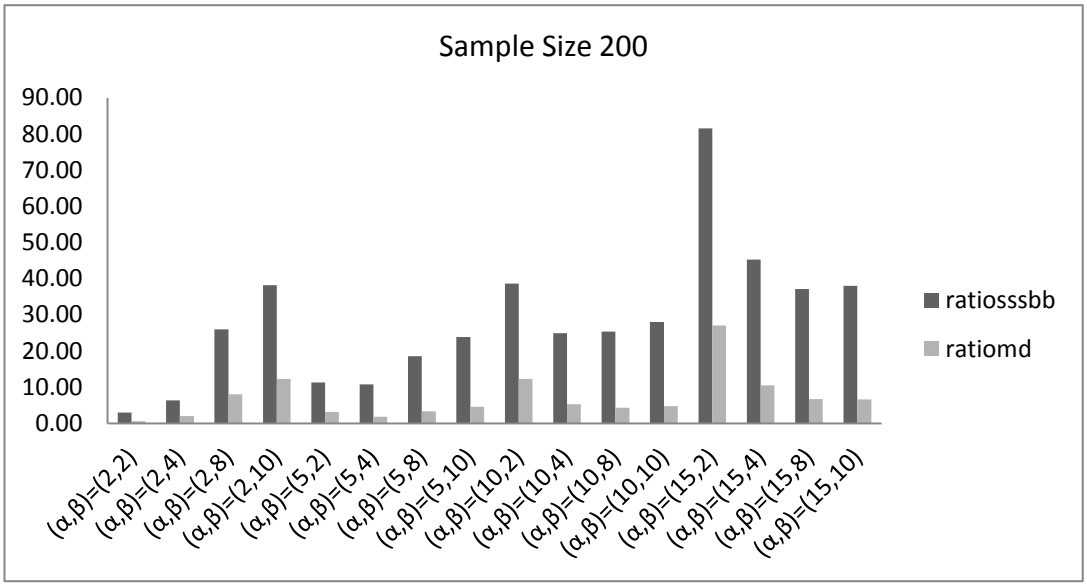


Table (4.4) shows the simulation results of both the techniques: SSSBB and Mahalanobis distance, done in MATLAB, outliers detected and the area of fence by both the techniques are calculated. As the table shows, SSSBB detect the possible outliers under the specified area and Mahalanobis distance covering more area detect more outliers. Row 1 of the table shows that with $\alpha=2$ and $\beta=2$ and sample size=40, the SSSBB ratio is 1.27 and the Mahalanobis distance ratio is 0.22, therefore the SSSBB performs better than the Mahalanobis distance. Row 2 indicates that with $\alpha=2$ and $\beta=4$, the average outliers detected by SSSBB are 1.35 and the area is 0.6, with ratio 2.61 while Mahalanobis distance detects average 2.78 outliers with area 4.91, so the ratio calculated

is 0.61. Therefore, the SSSBB is a good method for detecting outliers in beta distribution for different values of the parameters.

Figure 4.4(a) shows the graph of ratio of SSSBB and Mahalanobis distance for sample size 40, as it is clear from the graph that the ratio of SSSBB is greater than the Mahalanobis distance for different values of the parameters. For sample size 80, figure 4.4(b) shows the graph that the ratio of SSSBB is more than the Mahalanobis distance which means SSSBB detects the possible outlier in less area while Mahalanobis distance has fewer ratios. Figure 4.4(c) illustrates the graph of ratio of SSSBB and Mahalanobis distance, for beta distribution using different parameters values, shows that SSSBB ratio is greater than the Mahalanobis distance therefore SSSBB is a preferable method as compared to Mahalanobis distance.

4.5: REAL DATA

4.5.1 Pakistan's Stock Exchange

For empirical evidence, SSSBB and Mahalanobis distance are applied on the monthly data of Pakistan's stock exchange, for the detection of outliers. Stock exchange data is left skewed; the technique whose ratio of outlier detected and area of fence is greater will be a preferable method. The companies are selected on the basis of market capitalization, as we are considering bivariate data therefore, close and turn over data points are taken

and both the techniques are applied on the companies' data in order to check the performance of the techniques and get the superior technique.

Table 4.5 Performance of SSSBB and Mahalanobis distance on Pakistan's stock exchange

COMPANY	**SSSBB**			**MAHALANOBIS DISTANCE**		
	OD	AREA	RATIO	OD	AREA	RATIO
PPL	10.00	0.51	19.60	11.00	0.63	17.40
UBL	9.00	0.67	13.45	14.00	1.08	12.93
LUCK	14.00	1.77	7.92	14.00	5.30	2.64
ENGRO	9.00	2.56	3.52	9.00	5.27	1.71
POL	13.00	0.70	18.64	29.00	1.57	18.51

The table 4.5 shows the SSSBB and Mahalanobis distance applied on the data of the companies and SSSBB shows better result as compared to Mahalanobis distance because SSSBB ratio is more which means it detects possible outliers while Mahalanobis distance detect more outliers and covers more area. The ratio of companies PPL, UBL, LUCK, ENGRO and POL considered in the estimations is more in case of SSSBB that shows SSSBB detects possible outliers within the specified area of fence while less ratio of Mahalanobis distance shows that more outliers are detected by increasing the area of fence.

Figure 4.5 Performance of SSSBB and Mahalanobis distance on Pakistan's Stock Exchange data

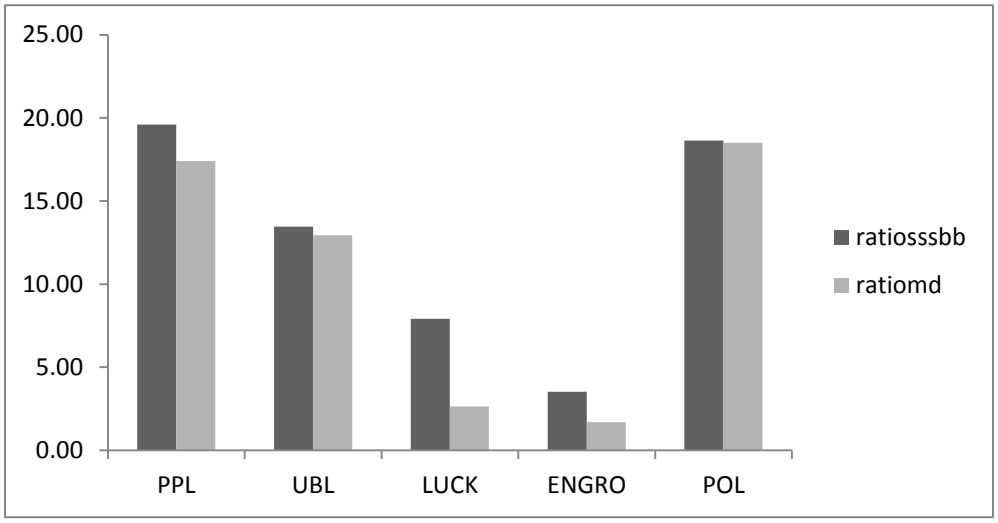


Figure 4.5 shows that SSSBB performs well as compared to Mahalanobis distance on the stock exchange data of Pakistan for the selected companies. The figure 4.5 shows that performance of SSSBB is better than Mahalanobis distance as the ratio of outlier detected and area of fence of SSSBB is more than the ratio of Mahalanobis distance.

4.5.2 Measures of Interest Rate

For bivariate data, two measures of interest rate that are money market rate and treasury bill rate data is used to check the performance of SSSBB and comparing its results with Mahalanobis distance. The method whose ratio of outlier detected and the area of fence are greater than the other method is considered to be better one. Following countries data are taken: Switzerland, United Kingdom, United States, Iceland, Spain and New Zealand.

Table 4.6 Performance of SSSBB and Mahalanobis distance on Measures of interest rate

Country	OD SSSBB	Area SSSBB	Ratio SSSBB	OD MD	Area MD	Ratio MD
Switzerland	33	14.87	2.21	71	63.57	1.11
UK	54	22.27	2.424	57	156.01	0.36
United States	39	6.12	6.36	70	58.99	1.18
Iceland	36	86.93	0.41	74	408.06	0.18
Spain	26	79.74	0.32	77	354.55	0.21
New Zealand	26	20.83	1.24	47	142.32	0.33

Table 4.6 shows the results of SSSBB and Mahalanobis distance applied on the two measures of interest rates: money market rates and Treasury bill rates as the performance of SSSBB and Mahalanobis distance is checked on bivariate data. The estimation results shows that SSSBB has greater ratio of outlier detected and the area of fence for the measures of interest rate than the Mahalanobis distance for all the countries considered in the study. The greater ratio of SSSBB means that the possible outliers are detected by SSSBB under the specified area of fence as compared to Mahalanobis distance that detect more outliers by increasing the area of fence. Therefore, SSSBB is a good method as compared to Mahalanobis distance.

Figure 4.6 Performance of SSSBB and Mahalanobis distance on Measures of interest rate

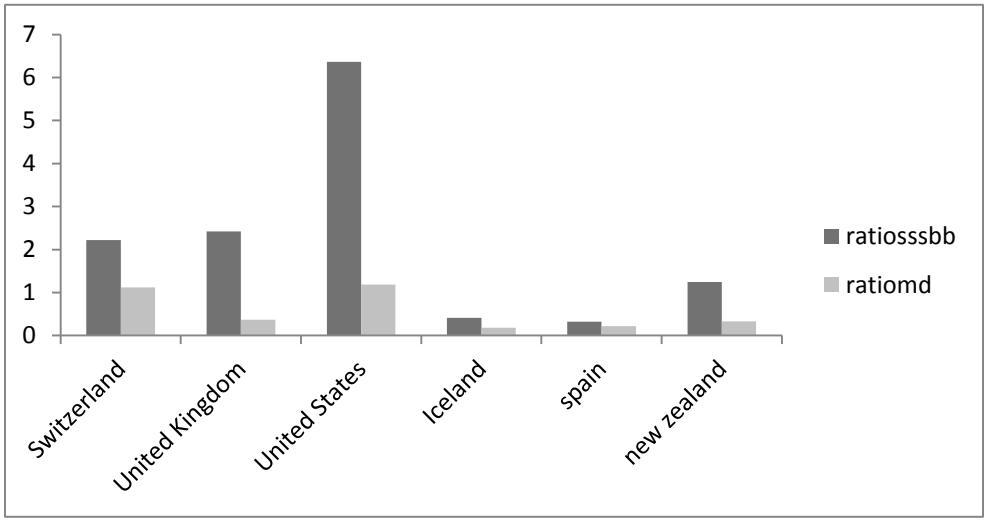


Figure 4.6 shows the graph of SSSBB and Mahalanobis distance for the measures of interest rate for the six countries considered in the study. As the graph shows the ratio of SSSBB is greater than the Mahalanobis distance that means SSSBB is a better method on the basis of ratio of outlier detected and the area of fence as compared to Mahalanobis distance.

CHAPTER 5

5. SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1: SUMMARY

Data analysis requires the first step of checking the data that whether it is appropriate for the analysis or it contains outliers which are important to detect, as it is an important unavoidable problem and give misleading results in the analysis. Thus, in order to have the appropriate results, it is important to manage the data accurately by identifying the outliers and treat them properly. There are two extreme choices in the analysis of outliers, whether to delete them with the risk of loss of information or to keep them, with the risk of contamination. Many methods are described in the literature for the detection of outliers and how to treat them but most of the methods are applicable, when there is symmetric data. The problem arises when the data is asymmetric because symmetry is not fulfilled everywhere, as most of the real data does not follow symmetric distribution. SSSBB is the method which considered symmetric as well as asymmetric data and proved its performance better than the existing ones by comparing the constructed fences with the true upper and lower critical values. SSSBB was available for the univariate data only.

But outliers are not easy to detect as the dimension and the number of outlier increases, as they can extend in multiple directions, when the multivariate data is considered. There were large number of methods proposed for the detection of outliers in

a univariate case but there are limited methods to detect outliers with the increase in the dimension of the data. SSSBB has been proposed in our study to detect the outliers in bivariate case and to check its performance; the proposed technique is compared with the robust Mahalanobis distance in the previous chapters, we did simulations in MATLAB. SSSBB and Mahalanobis distance are applied on symmetric and skewed distributions and on real data set in order to check the performance of the techniques.

As described in chapter 3, SSSBB used the orthogonal projections, to compute the horizontal and vertical distances of the data points from major axis and minor axis, respectively. The data is divided from the median to calculate the upper critical value and lower critical value using first and third quartiles and interquartile ranges. The observations which lie outside the interval are labeled as outliers. The ratio of outlier detected and the area of fence is used to check the performance of SSSBB and Mahalanobis distance on different distributions like t-distribution, chi-square distribution, gamma distribution and beta distribution.

5.2: CONCLUSION

The results show that SSSBB demonstrates the high level success in identifying outliers as compared to Mahalanobis distance in both the symmetric and skewed distributions, considered in the study. SSSBB performance is better for small as well as large data sets. For empirical evidence both the techniques are applied on bivariate data of Pakistan's stock exchange and on the interest rate measures that are Money market rates and Treasury bill rates, SSSBB performs well in all the data sets considered.

5.3: RECOMMNDATIONS

When a researcher is interested to detect outliers in a bivariate skewed data, one should use SSSBB instead of Mahalanobis distance as it is more simple and easy to understand. SSSBB is uniformly superior in both the symmetric and skewed distributions considered. SSSBB can be extended to trivariate and multivariate case to detect the possible outliers in the data set.

REFERENCES:

1. Acuna, E., & Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*.
2. Adil, I. H., & Irshad, A. U. R. (2015). A Modified Approach for Detection of Outliers. *Pakistan Journal of Statistics and Operation Research*, *11*(1).
3. Al-Shameri, K. S. S. (2014). *Robust Detection of Outlines in Univariate Data with Skewed Distribution* (Doctoral dissertation, Sudan University of Science and Technology).
4. Anscombe, F. J. (1960). Rejection of outliers. *Technometrics, 2*(2), 123-146.
5. Atkinson, A. C., & Riani, M. (1997). Bivariate boxplots, multiple outliers, multivariate transformations and discriminant analysis: the 1997 Hunter lecture. *Environmetrics*, *8*(6), 583-602.
6. Banerjee, S., & Iglewicz, B. (2007). A Simple Univariate Outlier Identification Procedure Designed for Large Samples. *Communications in Statistics- Simulation and Computation, 36*, 249-263.
7. Barnett, V. (1978). The Study of Outliers: Purpose and Model. *Applied Statistics , 27* (3), 242-250.
8. Barnett, V., & Lewis, T. (1984). *Outliers in statistical data* (3rd ed.). Wiley.
9. Beckman, R. J., & Cook, R. D. (1983). Outlier.....s. *Technometrics , 25* (2).

10. Billor, N., Hadi, A. S., & Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, *34*(3), 279-298.
11. Butler, R. W., Davies, P. L., & Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 1385-1400.
12. Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied statistics*, 231-237.
13. Chauvenet, W. (1876). *A treatise on plane and spherical trigonometry*. JB Lippincott & Company.
14. Cousineau, D., & Chartier, S. (2010). *Outlier Detection and Treatment;a review. International Journal of Pscyclogical Research* , 3 (1), 59-68.
15. Dixon, W. J. (1953). Processing data for outliers. *Biometrics*, *9*(1), 74-89.
16. Goldberg, K. M., & Iglewicz, B. (1992). Bivariate extensions of the boxplot. *Technometrics*, *34*(3), 307-320.
17. Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 27-58.
18. Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 761-771.
19. Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.
20. Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*(2), 85-126.

21. Holland, P. W., & Welsch, R. E. (1977). *Robust regression using iteratively reweighted least-squares. Communications in Statistics-theory and Methods, 6(9), 813-827.*
22. Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. *Wiley interdisciplinary reviews: Computational statistics*, *2*(1), 36-43.
23. Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, *52*(12), 5186-5201.
24. Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers* (Vol. 16). Asq Press.
25. Irwin, J, 0, 1925, *On a criterion for the rejection of outlying observations. Biometrika*, 17:238-250
26. Ludbrook, J. (2008). Outlying observations and missing values: how should they be handled? *Clinical and Experimental Pharmacology and Physiology*, *35*(5-6), 670-678.
27. McKay, A. T. (1935). The distribution of the difference between the extreme observation and the sample mean in samples of n from a normal universe. *Biometrika*, *27*(3/4), 466-471.
28. Mansur, M. O., Sap, M., & Noor, M. (2005, May). Outlier detection technique in data mining: a research perspective. In *Postgraduate Annual Research Seminar*.
29. Maronna, R. A., & Yohai, V. J. (1976). Robust estimation of multivariate location and scatter. *Wiley StatsRef: Statistics Reference Online*.

30. Nair, K. R. (1948). The distribution of the extreme deviate from the sample mean and its studentized form. *Biometrika*, *35*(1/2), 118-144.
31. Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical assessment, research & evaluation*, *9*(6), 1-12.
32. Peirce, B. (1852). Criterion for the rejection of doubtful observations. *The Astronomical Journal*, *2*, 161-163.
33. Pearson, K. (1931). Tables for statisticians and biometricians.
34. Peña, D., & Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, *43*(3), 286-310.
35. Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of statistics*, 1327-1345.
36. Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, *91*(435), 1047-1061.
37. Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212-223.
38. Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589). John wiley & sons.
39. Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, *8*, 283-297.
40. Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, *53*(4), 382-387.

41. Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis* (pp. 256-272). Springer US.
42. Sajesh, T. A., & Srinivasan, M. R. (2013). An Overview of Multiple Outliers in Multidimensional Data. *Sri Lankan Journal of Applied Statistics*, *14*(2).
43. Stone, E. J. (1868). On the rejection of discordant observations. *Monthly Notices of the Royal Astronomical Society*, *28*, 165-168.
44. Thompson, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics*, *6*(4), 214-219.
45. Tippett, L. H. (1925). On the extreme individuals and the range of samples taken from a normal population. *Biometrika*, *17*(3-4), 364-387.
46. Tongkumchum, P. (2005). Two-dimensional box plot. *Songklanakarin Journal of Science and Technology*, *27*(4), 859-866.
47. Tsay, R. S., Peña, D., & Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika*, *87*(4), 789-804.
48. Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesely.
49. Werner, M. (2003). *Identification of multivariate outliers in large data sets* (Doctoral dissertation, University of Colorado at Denver).
50. Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhyā: The Indian Journal of Statistics, Series A*, 407-426
51. Zaman, A., Rousseeuw, P. J., & Orhan, M. (2001). Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, *71*(1), 1-8.

52. Zani, S., Riani, M., & Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, *28*(3), 257-270.

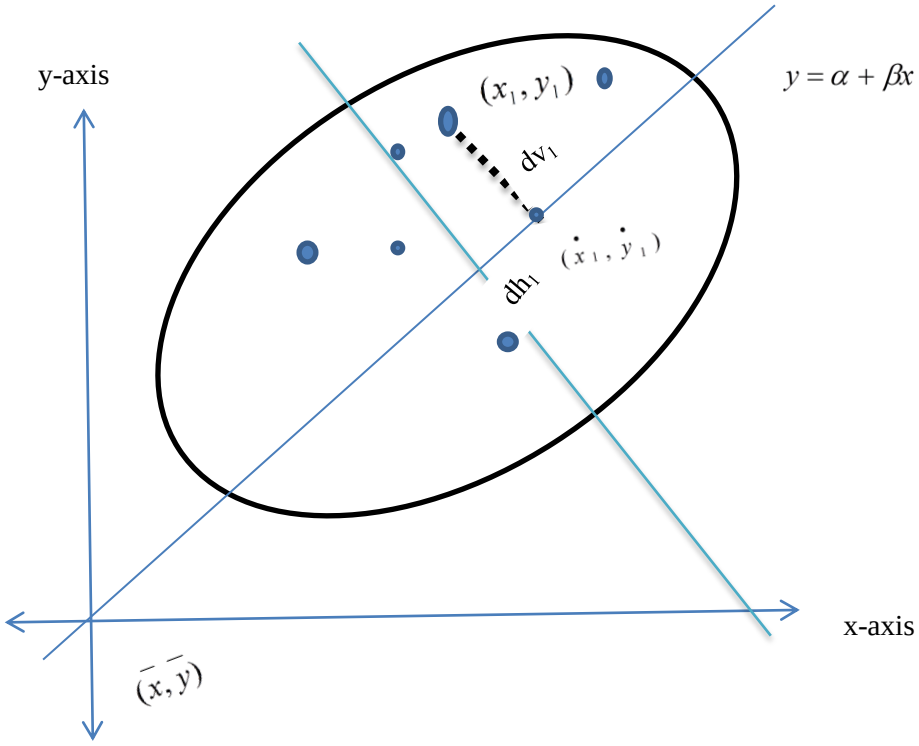
APPENDIX

Let us suppose a robust regression

$$Y = \alpha + \beta X + \in \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

For bivariate data, consider two variables x and y and suppose the points (x_1, y_1), (x_2, y_2) …

(x_n, y_n) are scattered data points of the variables.



Suppose a point (x_1, y_1) in the xy-axis.

The equation of the line is:

$(\dot{x}_1, \dot{y}_1)$ 1

Here m is the slope and c is the y-intercept.

The line perpendicular to this line is mathematically written as:

$$y = -\frac{1}{\beta}x + c \quad \text{..... 2}$$

For point (x_1, y_1), the equation of line is

$$y_1 = -\frac{1}{\beta}x_1 + c$$

Therefore, $c = y_1 + \frac{1}{\beta}x_1$ put the value of c in equation (2)

$$y = -\frac{1}{\beta}x + (y_1 + \frac{1}{\beta}x_1) \quad \text{..... 3}$$

Using equation 1 and 3 to find the point $(\dot{x}_1, \dot{y}_1)$,

$$y_1 + \frac{1}{\beta}x_1 - \alpha = \beta\dot{x}_1 + \frac{1}{\beta}\dot{x}_1 \quad \text{..... 4}$$

To find the point $\dot{x}_1$, equation 4 can be written as:

$$\dot{x}_1 = (\beta + \frac{1}{\beta})^{-1}(y_1 + \frac{1}{\beta}x_1 - \alpha)$$

Point $\dot{y}_1$ can be found as:

$$\dot{y}_1 = \alpha + \beta \dot{x}_1$$

Vertical distance, d_v, will be calculated using distance formula as:

$$d_v = \sqrt{(x_1 - \dot{x}_1)^2 + (y_1 - \dot{y}_1)^2} \quad \text{..... 5}$$

And the horizontal distance, d_h,

$$d_h = \sqrt{(\dot{x}_1 - \bar{x})^2 + (\dot{y}_1 - \bar{y})^2} \quad \text{..... 6}$$

Similarly for all the data points, horizontal and vertical distances are calculated using equation 5 and equation 6.