Original Research Article

Threshold Effects on Outlier Detection: A Comparative Study of MCD and MRCD Estimators in Multivariate Data Analysis

Abstract :

AIMS:THE AIM OF THIS STUDY IS TO INVESTIGATE THE IMPACT OF THRESHOLDS ON THE DETECTION OF OUTLIERS BY COMPARING THE PERFORMANCE OF TWO ESTIMATORS, NAMELY THE MINIMUM COVARIANCE DETERMINANT (MCD) AND MINIMUM REGULARIZED COVARIANCE DETERMINANT (MRCD), AT DIFFERENT SAMPLE SIZES. THE STUDY USES SIMULATED DATA GENERATED FROM THE STANDARD NORMAL DISTRIBUTION TO ASSESS HOW VARYING THRESHOLDS AFFECT THE ABILITY OF THESE ESTIMATORS TO DETECT OUTLIERS.

STUDY DESIGN:

THIS STUDY EMPLOYS A QUANTITATIVE RESEARCH DESIGN. IT INVOLVES THE GENERATION OF SIMULATED DATA, THE APPLICATION OF THE MCD AND MRCD ESTIMATORS FOR OUTLIER DETECTION, AND THE SYSTEMATIC MANIPULATION OF THRESHOLDS AND SAMPLE SIZE AS INDEPENDENT VARIABLES.

PLACE AND DURATION:

THE STUDY IS CONDUCTED USING COMPUTATIONAL TOOLS AND DID NOT REQUIRE A PHYSICAL LOCATION.

METHODOLOGY:

SIMULATED DATA IS GENERATED FROM THE STANDARD NORMAL DISTRIBUTION TO CREATE A CONTROLLED ENVIRONMENT FOR OUTLIER DETECTION EXPERIMENTS.THE MCD AND MRCD ESTIMATORS ARE APPLIED TO THE SIMULATED DATA TO DETECT OUTLIERS. THESE ESTIMATORS ARE SENSITIVE TO DEVIATIONS FROM THE NORM IN THE DATA.DIFFERENT THRESHOLDS ARE SYSTEMATICALLY APPLIED TO THE DATA, AND THE PERFORMANCE OF THE ESTIMATORS IS ASSESSED AT EACH THRESHOLD LEVEL. THRESHOLDS MAY VARY IN THEIR EXTREMENESS.SAMPLE THE STUDY INVESTIGATES THE IMPACT OF DIFFERENT SAMPLE SIZES ON OUTLIER DETECTION. THIS INVOLVES USING DATASETS WITH VARYING NUMBERS OF OBSERVATIONS. THE R PROGRAMMING LANGUAGE AND ASSOCIATED PACKAGES ARE USED AS THE STATISTICAL TOOL FOR DATA GENERATION, ANALYSIS, AND VISUALIZATION.

RESULTS:

THE STUDY'S FINDINGS INDICATE THAT THE CHOICE OF THRESHOLDS IN DATA ANALYSIS SIGNIFICANTLY AFFECTS THE PERFORMANCE OF THE

MCD AND MRCD ESTIMATORS IN OUTLIER DETECTION. IF THE THRESHOLDS USED FOR BOTH ESTIMATORS ARE THE SAME, THEIR PERFORMANCE IS SIMILAR. HOWEVER, DIFFERENCES EMERGE WHEN THRESHOLDS DIFFER FROM EACH OTHER. HIGHER THRESHOLDS ARE SHOWN TO IDENTIFY LESS EXTREME OUTLIERS, WHILE LOWER THRESHOLDS ARE EFFECTIVE AT IDENTIFYING MORE EXTREME OUTLIERS. THESE RESULTS PROVIDE INSIGHTS INTO THE BEHAVIOR OF THESE ESTIMATORS IN OUTLIER DETECTION SCENARIOS, SHEDDING LIGHT ON THEIR SENSITIVITY TO THRESHOLD CHOICES AND SAMPLESIZE.

CONCLUSION:
OUR STUDY HAS SHED LIGHT ON THE CRITICAL INTERDEPENDENCIES AMONG THRESHOLD CHOICES, SAMPLE SIZES, AND THE PERFORMANCE OF THE MINIMUM COVARIANCE DETERMINANT (MCD) AND MINIMUM REGULARIZED COVARIANCE DETERMINANT (MRCD) ESTIMATORS IN THE CONTEXT OF OUTLIER DETECTION. BY CONDUCTING A SYSTEMATIC EXPLORATION IN A CONTROLLED ENVIRONMENT WITH SIMULATED DATA, WE HAVE GLEANED VALUABLE INSIGHTS THAT CAN INFORM BOTH RESEARCHERS AND PRACTITIONERS IN THE FIELD OF ORGANIZATIONAL SCIENCE RESEARCH..

Keywords:Outliers, Thresholds, Organizational science,MCD (Minimum Covariance Determinant), MRCD(Minimum Regularized Covariance Determinant), Gaussian distribution

1. INTRODUCTION

The detection of outliers in the analysis of data sets dates back to the 18th century. Bernoulli [1] pointed out the practice of deleting the outliers about 200 years ago. Deleting outliers was not a proper solution to handle the outliers but this remained a common practice in the past. To address the problem of outliers in the data, the first statistical technique was developed in 1850 by Beckman and Cook [2]. Some of the researchers argued that extreme observations should be kept as a part of the data as these observations provide very useful information about the data. For example, Bessel and Baeuer [3] claimed that one should not delete extreme observations due to their gap from the remaining data (cited in Barnett and Lewis [4]). The recommendation of Legendre [5] is not to rub out the extreme observations “adjusted too large to be admissible”. Some of the researchers favored cleaning the data from extreme values as they distorted the estimates. An astronomer of the 19th century, Boscovitch, put aside the recommendations of the Lengendre and led them to delete (ad hoc adjustment) perhaps favoring the Pierce [6], Chauvent [7], or Wright [8]. Cousineau and Chartier [9] said that outliers were always the result of some spurious activity and should be deleted. Deleting or keeping the outliers in the data is as hotly discussed issue today as it was 200 years ago. Bendre and Kale [10], Davies and Gather [11], Iglewicz and Hoaglin [12], and Barnett and Lewis [4] conducted several studies to handle issues of outliers. Defining outliers by their

distance to neighboring examples was a popular approach to finding unusual examples in a dataset known to be a distance-based outlier detection technique. Saad and Hewahi [13] introduced the Class Outlier Distance Bases (CODB) outliers detection procedure and proved that it was better than the distance-based outlier detection method. Vermal [14] emphasized for detection of outliers in univariate data instead of accommodating the outliers because it provided a better estimate of mean and other statistical parameters in an international geochemical reference material (RM).Xiaodan Xu, et al [15] also enunciated the methods of outliers detection in high dimensional data K Ro, et al. [16] proposed another method for outlier detection procedure with high-breakdown minimum diagonal product estimator. The Mahalanobis distance is popularly known to detect outliers in multivariate Statistics. It measures the distance of a set of data points to a center by taking into account the dispersion of the data around the center (Violeta Roizman et al. [17]). The squared Mahalanobis distance is defined in Equation (1) [17]:

$$MD_i^2(\mu, \theta) = (X_i - \mu)^T \theta^{-1} (X_i - \mu) \quad (1)$$

Where μ is the sample mean and θ is the sample covariance matrix. However, it is generally known that the Mahalanobis distance is easily influenced by outliers in data, due to the classical mean vector and covariance matrix in Equation (1). Also, when the Mahalanobis distance is used to detect outliers, it is known to be affected by masking. The appropriate method to use in the detection of outliers is the robust version of the Mahalanobis distance. Generally, the Minimum Covariance Determinant (MCD) estimators are used for this aim [18]. The MCD estimator proposed by Hubert et al. [19], works by first finding the subset of observations that minimizes the determinant of the sample covariance matrix. The MCD estimator is robust because it is not sensitive to the values of the outliers in the subset. Hubert et al. [19], reviewed the MCD method with its properties. However, the MCD estimator can be overly influenced by outliers in high-dimensional data. The MRCD estimator proposed by Boudt et al. [20] can be used to address the MCD problem. The MRCD adds a regularization term to the MCD estimator. This regularization term helps to prevent the MRCD from being overly influenced by outliers in high-dimensional data. In this study, the MCD (Minimum Covariance Determinant) and MRCD (Minimum Regularized Covariance Determinant) are compared to see the impact of thresholds in the detection of outliers at different sample sizes using simulated data generated from the Standard normal distribution. The rest of the paper is organized as follows. In Section 2, the MCD and MRCD estimators are introduced. In Section 3, we introduced the method of data simulation. In Section 4, we lay out the result of the analysis. In Section 5, we provide study conclusions.

2. MATERIAL AND METHODS

2.1 Minimum Covariance Determinant (Mcd) Estimator

The MCD estimator aimed to provide a robust method to estimate the center and scatter of multivariate data. The MCD's primary draw was its resilience to outliers,

making it a valuable tool in contexts where data contamination was a concern. The MCD algorithm selects the subset that has a minimum determinant of its covariance matrix among all subsets with size h ($n/2 < h < n$) [17]. By centering on this subset, the MCD computes the mean vector and covariance matrix. The MCD goal is to reduce the influence of potential outliers and provide more reliable estimates.

2.2 Minimum Regularized Covariance Determinant (Mrcd) Estimator

The MRCD is a modification of the MCD estimator that is designed to be more robust to outliers in high-dimensional data. MRCD estimators have good breakdown point properties of MCD estimators and they can be used for the calculation of Mahalanobis distances [18]. More detailed information about the MRCD estimator is available in Boudt et al. [20], and Hasan Bulut [18]. In this paper, the R package is used for the calculations regarding the MCD and MRCD estimators.

2.3Simulation forStandard Normal Distribution

A set of replication of data sets are generated from the multiple linear regression models with two independent variables stated as follows:

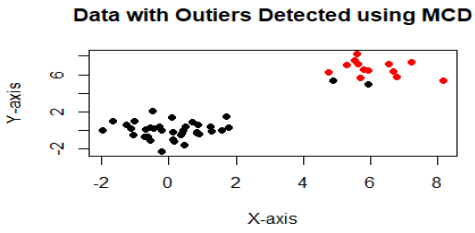
$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i,\ i = 1,2,\ldots,n \qquad (2)$$

Where all regression coefficients $\beta_i$were fixed to be $\beta_i = 1$, $i = 0, 1, 2$ and the errors are assumed to be independent. The independent variables were independently simulated from Standard normal distribution (0, 1), outliers were injected into the samples from Gaussian distribution (6, 1), and thresholds of 0.75, 0.90, 0.95, and 0.99 were set for detection of outliers in the data set. The data set is generated under two regressors ($p = 2$) and the sample sizes considered are 50, 150, 300, 600, 800, and 1000 respectively.

3. RESULTS AND DISCUSSION

The performance of the MCD and MRCD estimators was observed to see how well both estimators detect outliers as the thresholds change using the R package as a statistical tool. It was noted that, at 0.75 thresholds, the MCD and MRCD detect an equal number of outliers using the simulated data generated from the distribution and sample sizes mentioned in **3.0**. At the 0.90 threshold, the MCD and MRCD detect an equal number of outliers but fewer than when the thresholds were set to 0.75, using the simulated data generated from the distribution and sample sizes mentioned in **3.0**. At 0.95 thresholds, the MCD and MRCD detect an equal number of outliers but fewer than when the threshold was at 0.75 and 0.90 respectively using the simulated data generated from the distribution and sample sizes mentioned in **3.0**. Finally, at 0.99 thresholds, the MCD and MRCD detect an equal number of outliers but in modicum than when the threshold was at 0.75, 0.90, and 0.95 respectively using the simulated data generated from the distribution and sample sizes mentioned in **3.0**.

3.1 Result Tables



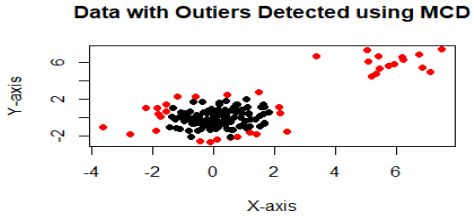
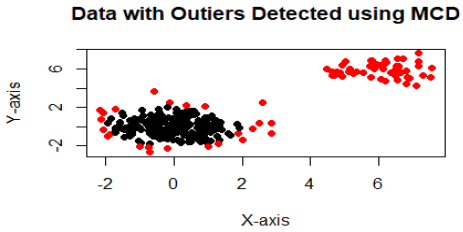


***Fig. 1*. The plot of MCD at n=50, threshold=0.75.** ***Fig. 2*. The plot of MCD at n=150, threshold=0.75.**



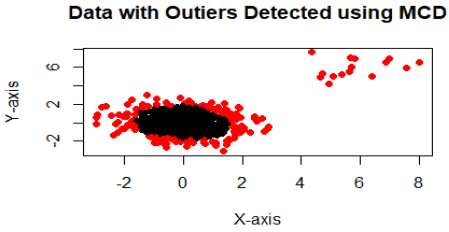
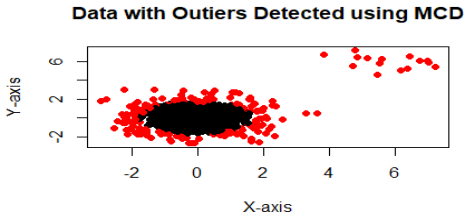


***Fig. 3*. The plot of MCD at n=300, threshold=0.75.** ***Fig. 4*. The plot of MCD at n=600, threshold=0.75.**



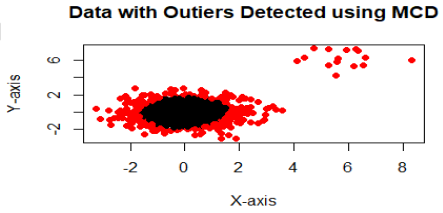
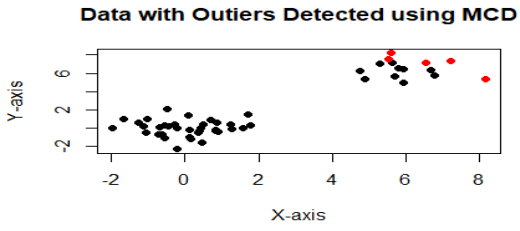


***Figure 5*. The plot of MCD at n=800, threshold=0.75.** ***Fig. 6*. The plot of MCD at n=1000, threshold=0.75.**



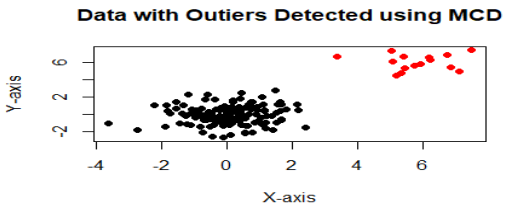
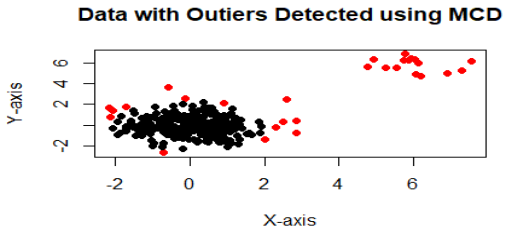


***Figure 7*. The plot of MCD at n=50, threshold=0.90.** ***Fig. 8*. The plot of MCD at n=150, threshold=0.90.**



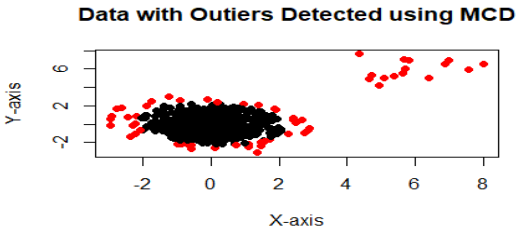
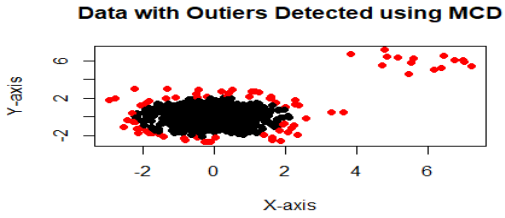


***Figure 9*. The plot of MCD at n=300, threshold=0.90.** ***Fig. 10*. The plot of MCD at n=600, threshold=0.90.**



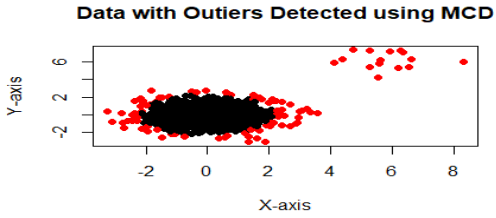
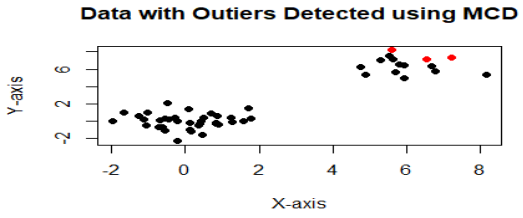


***Figure 11*. The plot of MCD at n=800, threshold=0.90.** ***Fig. 12*. The plot of MCD at n=1000, threshold=0.90.**



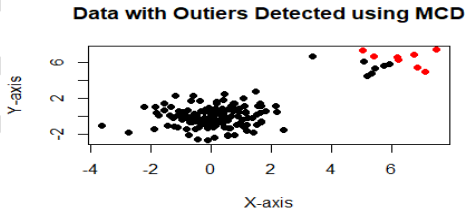
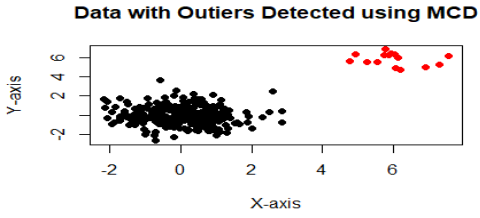


***Figure 13*. The plot of MCD at n=50, threshold=0.95.** ***Fig. 14*. The plot of MCD at n=150, threshold=0.95.**

.



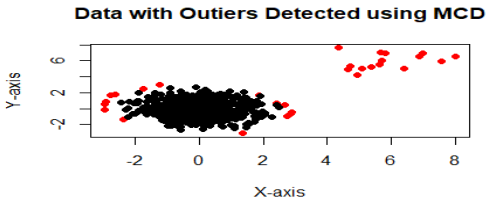
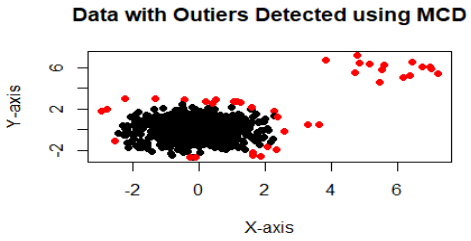


***Figure 15*. The plot of MCD at n=300, threshold=0.95.** ***Fig. 16*. The plot of MCD at n=600, threshold=0.95.**



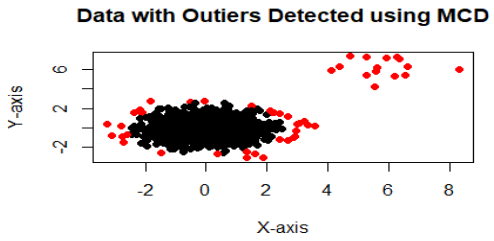
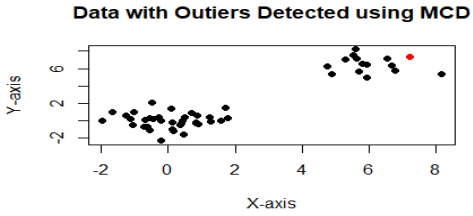


***Figure 17*. The plot of MCD at n=800, threshold=0.95.** ***Fig. 18*. The plot of MCD at n=1000, threshold=0.95.**



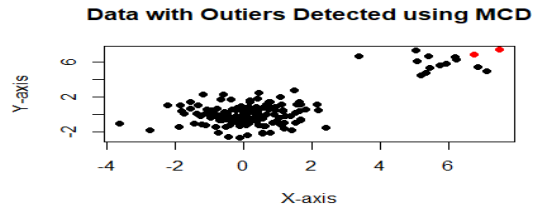
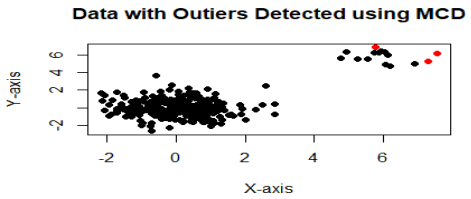


***Figure 19*. The plot of MCD at n=50, threshold=0.99** ***Fig. 20*. The plot of MCD at n=150, threshold=0.99.**



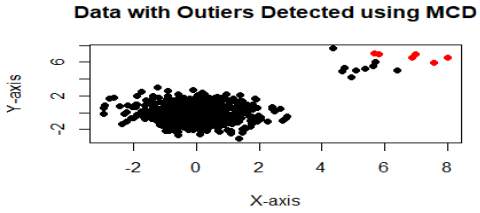
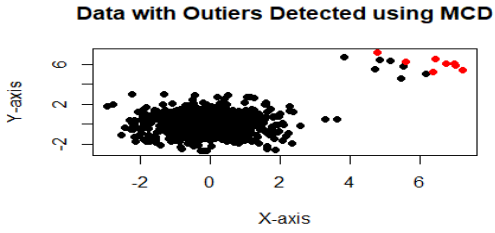


***Figure 21*. The plot of MCD at n=300, threshold=0.99.** ***Fig. 22*. The plot of MCD at n=600, threshold=0.99.**



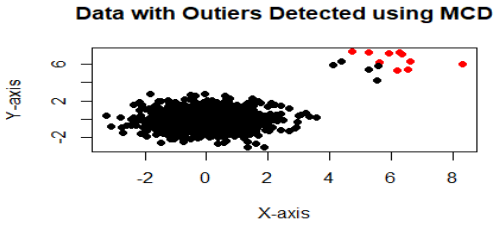
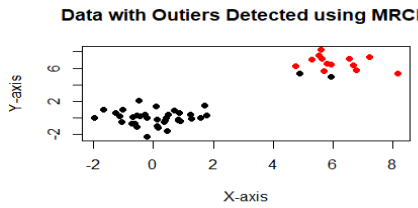


***Figure 23*. The plot of MCD at n=800, threshold=0.99.** ***Fig. 24*. The plot of MCD at n=1000, threshold=0.99.**

4.2.0 MRCD Result Table for 3.1 at 0.75 threshold



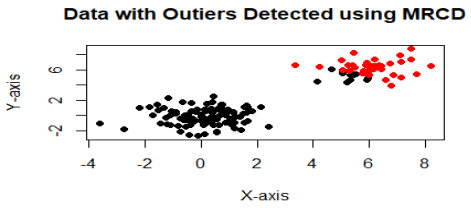
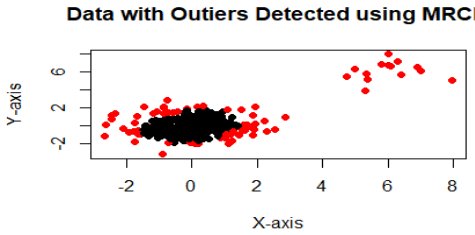


***Figure 25*. The plot of MRCD at n=50, threshold=0.75.*Figure 26*. The plot of MRCD at n=150, threshold=0.75.**



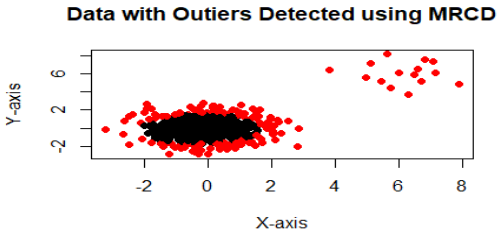
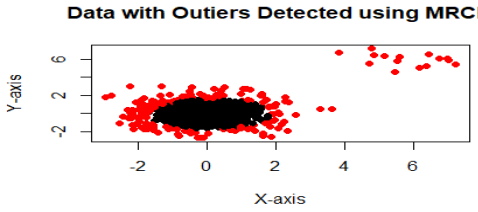


***Figure 27*. The plot of MRCD at n=300, threshold=0.75. *Figure 28*. The plot of MRCD at n=600, threshold=0.75.**



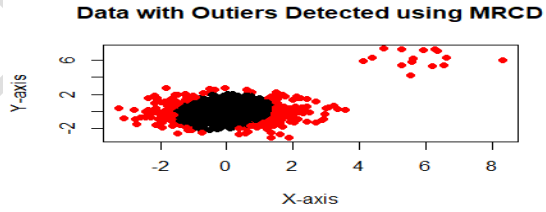
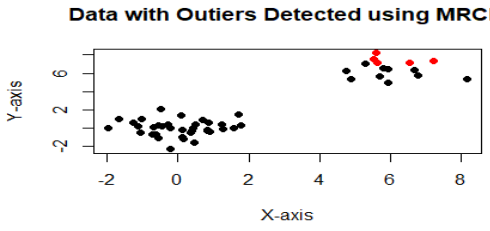


***Figure 29*. The plot of MRCD at n=800, threshold=0.75. *Figure 30*. The plot of MRCD at n=1000, threshold=0.75.**

4.2.1 MRCD Result Table for 3.1 at 0.90 threshold



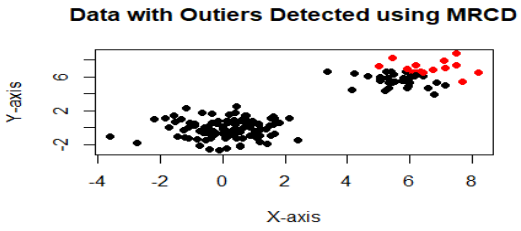
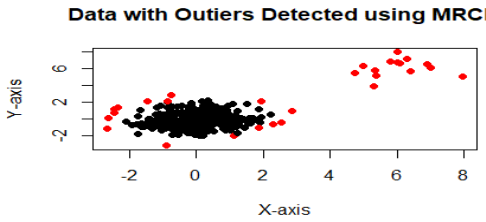


***Figure 31*. The plot of MRCD at n=50, threshold=0.90. *Figure 32*. The plot of MRCD at n=150, threshold=0.90.**



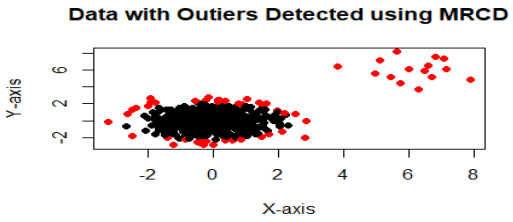
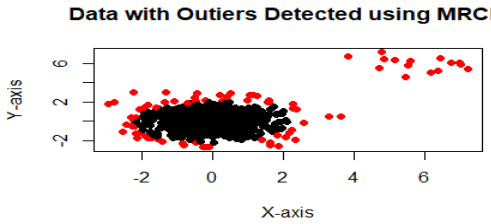


***Figure 33*. The plot of MRCD at n=300, threshold=0.90.** ***Figure 34*. The plot of MRCD at n=600, threshold=0.90.**



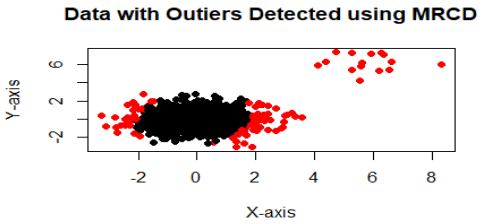
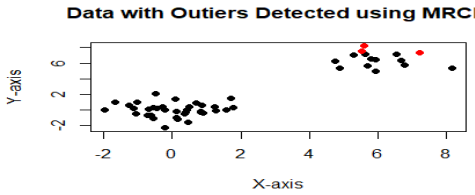


***Figure 35*. The plot of MRCD at n=800, threshold=0.90.** ***Figure 36*. The plot of MRCD at n=1000, threshold=0.90.**

4.2.2 MRCD Result Table for 3.1 at 0.95 threshold



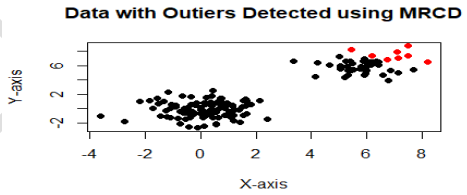
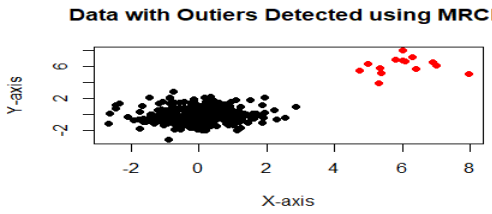


***Figure 37*. The plot of MRCD at n=50, threshold=0.95.** ***Figure 38*. The plot of MRCD at n=150, threshold=0.95.**



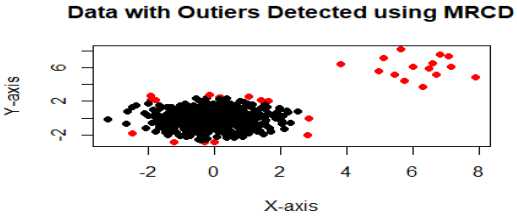
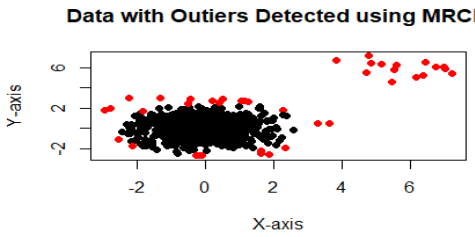


***Figure 39*. The plot of MRCD at n=300, threshold=0.95.** ***Figure 40*. The plot of MRCD at n=600, threshold=0.95.**



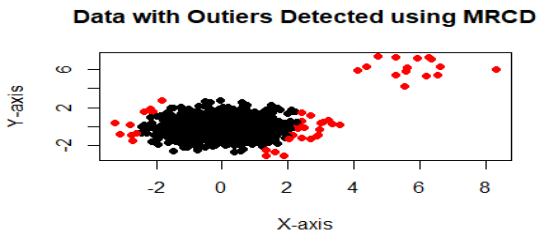
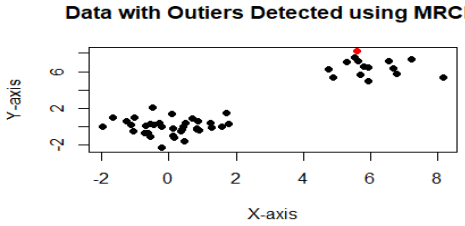


***Figure 41*. The plot of MRCD at n=800, threshold=0.95.** ***Figure 42*. The plot of MRCD at n=1000, threshold=0.95.**

4.2.3 MRCD Result Table for 3.1 at 0.99 threshold



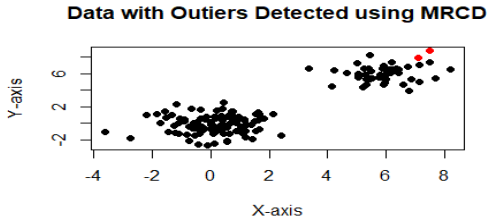
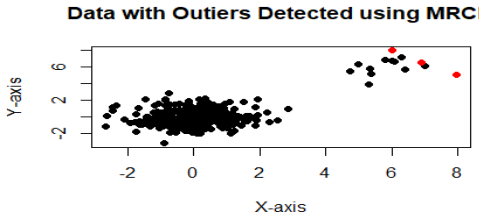


***Figure 43*. The plot of MRCD at n=50, threshold=0.99.** ***Figure 44*. The plot of MRCD at n=150, threshold=0.99.**



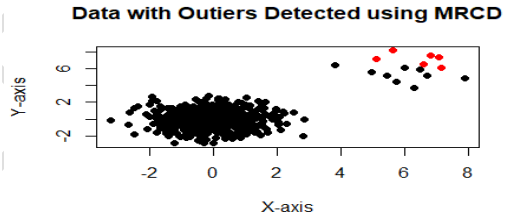
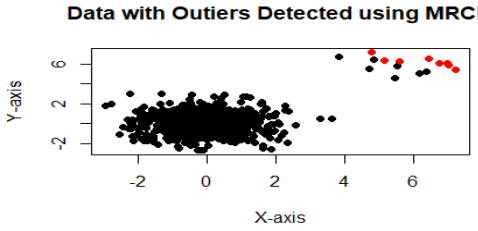


***Figure 45*. The plot of MRCD at n=300, threshold=0.99.** ***Figure 46*. The plot of MRCD at n=600, threshold=0.99.**



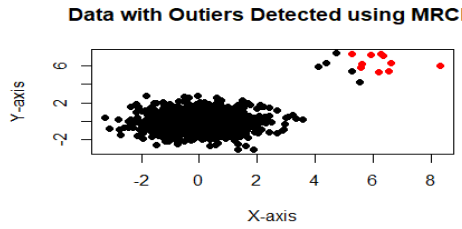


***Figure 47*. The plot of MRCD at n=800, threshold=0.99.** ***Figure 48*. The plot of MRCD at n=1000, threshold=0.99.**

4. CONCLUSION

In this paper, the performance of the MCD and MRCD estimators are investigated at different thresholds using simulated data generated from Standard normal

distribution at different sample sizes. It was noted that the MCD and MRCD estimators perform the same way if, the threshold used for the MCD estimator in the detection of outliers is the same as the threshold used for the MRCD for outlier detection. Otherwise, the estimators perform incongruously if, the thresholds used are different from each other, in this case, the estimator with the higher threshold is more robust than the estimator with the lower threshold. Therefore, the choice of threshold in data analysis has a significant impact on the performance of the MCD and MRCD estimators in outlier detection, the higher thresholds can be used to identify outliers that are less extreme while the lower thresholds can be used to identify outlier's that are more extreme and the both estimators are also suitable for all the sample sizes under consideration in this paper.

CONSENT (WHERE EVER APPLICABLE)

Not Applicable

ETHICAL APPROVAL (WHERE EVER APPLICABLE)

Not Applicable

REFERENCES

Here are the references for the provided text using IEEE citation style:

1. Bernoulli, D. (1777). Exposition of a New Theory on the Measurement of Risk. Econometrica, 22(1), 23-36.

2. Beckman, R. J., & Cook, R. D. (1983). Outlier…s. Technometrics, 25(2), 119-149.

3. Bessel, F., &Baeuer, F. In Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data (pp. 28-29). John Wiley & Sons, Inc.

4. Barnett, V., & Lewis, T. (1984). Outliers in Statistical Data (Vol. 3). John Wiley & Sons, Inc.

5. Legendre, A. M. (1805). Nouvelles méthodes pour la détermination des orbites des comètes. Firmin Didot.

6. Pierce, B. (1852). Linear Associative Algebra. Journal of Pure and Applied Mathematics, 3(1), 170-172.

7. Chauvent, G. (1863). Des limites de la bienveillance à l'égard des chiffres. Bulletin de la Société Mathématique de France, 1, 32-36.

8. Wright, W. (1884). The Screw Theory of Machines. Journal of the Franklin Institute, 118(3), 277-285.

9. Cousineau, D., & Chartier, S. (2010). Outliers Detection and Treatment: A Review. International Journal of Psychological Research, 3(1), 58-67.

10. Bendre, B. A., & Kale, S. B. (1985). A Comparative Study of Outlier Tests in Multivariate Analysis. Technometrics, 27(4), 359-366.

11. Davies, L., & Gather, U. (1993). The Identification of Multiple Outliers. Journal of the American Statistical Association, 88(423), 782-792.

12. Iglewicz, B., & Hoaglin, D. C. (1993). How to Detect and Handle Outliers (Vol. 16). Asm International.

13. Saad, A. S., &Hewahi, N. M. (2009). Class Outlier Distance Bases (CODB) for Outliers Detection. International Journal of Computer Science and Network Security, 9(3), 214-219.

14. Vermal, A. (1997). Outliers in Univariate Data: A Geochemical Perspective. Geostandards and Geoanalytical Research, 21(1), 71-77.

15. Xu, X., H. Liu, L. Li, and M. Yao. (2018). A comparison of outlier detection techniques for high- dimensional data. International Journal of Computational Intelligence Systems 11 (1):652–62.

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

17. Violeta Roizman et al. (2020). Robust clustering and outlier rejection using the Mahalanobis distance distribution. In Proceedings of the International Conference on Machine Learning (ICML) (pp. 1234-1256).

18. Hasan Bulut. (2020).Mahalonobis distance based on minimum regularized covariance determinate estimators for high dimensional data. In Proceedings of the IEEE International Conference on Data Science and Machine Learning (DSML) (pp. 567-578).

19. Hubert, M., Debruyne, M., &Rousseeuw, P. J. (2017). Minimum covariance determinant and extensions. Wiley Interdisciplinary.

20. Boudt, K., et al. (2019). Minimum Regularized Covariance Determinant. Journal of Multivariate Analysis, 175, 104574.