Open Journal of Statistics, 2019, 9, 15-26
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Outlier Detection Based on Robust Mahalanobis Distance and Its Application

Xu Li[1], Songren Deng[1], Lifang Li[1], Yunchuan Jiang[2*]

[1]School of Economic, Jinan University, Guangzhou, China
[2]Department of Anatomy, Division of Basic Medicine, YongZhou Vocational Technical College, Yongzhou, China
Email: *147750594@qq.com

How to cite this paper: Li, X., Deng, S.R., Li, L.F. and Jiang, Y.C. (2019) Outlier Detection Based on Robust Mahalanobis Distance and Its Application. *Open Journal of Statistics*, **9**, 15-26.
https://doi.org/10.4236/ojs.2019.91002

Received: December 28, 2018
Accepted: January 21, 2019
Published: January 24, 2019

Abstract

Classical Mahalanobis distance is used as a method of detecting outliers, and is affected by outliers. Some robust Mahalanobis distance is proposed via the fast MCD estimator. However, the bias of the MCD estimator increases significantly as the dimension increases. In this paper, we propose the improved Mahalanobis distance based on a more robust Rocke estimator under high-dimensional data. The results of numerical simulation and empirical analysis show that our proposed method can better detect the outliers in the data than the above two methods when there are outliers in the data and the dimensions of data are very high.

Keywords

MCD Estimator, Rocke Estimator, Outlier, Mahalanobis Distance

1. Introduction

With the advancement of information technology, all fields have gradually entered into the era of big data. In addition, with more comprehensive research data in various research fields, it also brings trouble to the processing data. In the case of more variables that need to be detected, collected and processed, the greater the probability of errors will cause, the more the number of outliers in the data will increase. In fact, the data will be affected by various complicated and uncertain factors, as well as the occurrence of outliers due to the accuracy of the instrument, statistical omissions, and operational errors. Therefore, detecting outliers are generally required before analyzing data. Due to the increase of the probability of occurrence of outliers in high-dimensional data, it is more necessary to detect outliers in high-dimensional data. For one-dimensional data, there are many methods for determining outliers, e.g., three standard deviation standards, box plots [1], etc. However, in high-dimensional data, because some variables may have a certain correlation, the exception of one of the variables does not indicate that this is an outlier, so the method of detecting outliers for one-dimensional data is not directly applicable to high-dimensional data.

There are many methods for detecting outliers in high-dimensional data. For example, leverage value [2] [3] [4], Mahalanobis distance [5], genetic algorithm [6]. Previous studies include: Yan *et al.* [7] proposed anomaly diagnosis based on leveraged large dataset sampling, Shi *et al.* [8] proposed a high-dimensional outlier detection algorithm based on genetic algorithm, and so on. The classical Mahalanobis distance is a common method for detecting outliers. However, it is a method based on sample mean vector and sample covariance matrix. Since the classical mean vector and covariance matrix algorithms are sensitive to outliers, the classical Mahalanobis distance is also sensitive to outliers. Many authors have proposed robust estimation methods for mean vector and covariance matrix, such as S-D estimator (Stahel (1981) [9], Donoho (1982) [10]), MVE estimator [11], MCD estimator (Grübel R, 1988) [12], S estimator (Rousseeuw and Yohai, 1983) [13], etc. At the same time, the robust Mahalanobis distance is proposed in the literature based on the robust mean vector and covariance matrix. In addition, the fast MCD estimator [14] is widely applied since its computation is simple and fast, and it has high robustness. In 2005, Wang [15] proposed a robust Mahalanobis distance based on fast MCD estimator. In this paper, the robust sample Mahalanobis distance is calculated based on the fast MCD estimator. Compared with the classical Mahalanobis distance, there is a good improvement in robustness. In 2014, Feng *et al.* [16] applied the robust Mahalanobis distance based on fast MCD estimator to the analysis of LiDAR point cloud data. In 2017, Maronna and Yohai [17] pointed out that the bias of the fast MCD estimator increased as the data dimension increased, and then proposed a Rocke estimator. The paper shown that with equal efficiencies, comparing to S-D estimator, MCD estimator and MM estimator etc, the Rocke estimator has the best robust when the data dimension was larger than 15. Also its computing time is competitive for data dimension was less than 100, and can presumably be improved. Due to the increasing number of variables in practical applications, and MCD estimator is not robust under high dimensional data, we need a robust Mahalanobis distance algorithm for high-dimensional data. Thus, in this paper, we propose a robust Mahalanobis distance algorithm based on the Rocke estimator. The numerical simulation and a real data analysis show that our proposed method can better detect the outliers in the data than the Mahalanobis distance method and the robust Mahalanobis distance base on the fast MCD estimator when there are outliers in the data and the dimensions of data are very high.

The rest of the paper is organized as follows. In Section 2, we introduce the principle and application of Mahalanobis distance and the basic principles and algorithms of Rocke estimator, and propose an algorithm for detecting the outlier of robust Mahalanobis distance in high-dimensional data. In Section 3, simulation studies are conducted to evaluate the finite sample performance of the proposed methods. In Section 4, a real data set is analyzed to compare the proposed methods with the existing methods. A discussion is given in Section 5.

2. Methodology and Algorithm

2.1. Principle of Mahalanobis Distance

The Mahalanobis distance was proposed by the Indian statistician Mahalanobis [5]. It represents a covariance distance of data, which can effectively estimate the similarity of sample sets. Compared with Euclidean distance, the Mahalanobis distance considers the correlation between features and is dimensionless. For a p-dimensional data $x=\left(x_1,x_2,x_3,\cdots,x_p\right)^{\mathrm{T}}$ with mean vector $\mu=\left(\mu_1,\mu_2,\mu_3,\cdots,\mu_p\right)^{\mathrm{T}}$ and covariance matrix Σ, the Mahalanobis distance is defined as follows:

$$D_M\left(x\right)=\sqrt{\left(x-\mu\right)^{\mathrm{T}}\Sigma^{-1}\left(x-\mu\right)} \tag{1}$$

It can be understood as the difference between the mean of x and the sample data, that is, the difference between x and the total sample position. If the difference is larger, it is considered that the possibility that x is not the total sample source is greater. In addition, when Σ is an identity matrix, the Mahalanobis distance is the same as the Euclidean distance.

2.2. Application of Mahalanobis Distance

In data mining, such as clustering, classification and other algorithms, the distance function is applied, and the Mahalanobis distance is one of the most commonly used distances. Furthermore, in the field of signal processing, information security, and even biomedicine, astronomy is inseparable from the concept of distance. Therefore, the Mahalanobis distance is very meaningful for these researches. For a set of samples $X_{n,p}$ with n and dimensions of p, we first calculate the mean vector μ and covariance matrix Σ of the sample $X_{n,p}$, and then calculate the Mahalanobis distance of each sample. We identify whether a point is an outlier, a threshold is needed. We know that $\sqrt{\left(x-\mu\right)^{\mathrm{T}}\Sigma\left(x-\mu\right)}$ approximates a chi-square distribution with a degree of p. Therefore, given a confidence level α, if there is $d>\chi_p^2\left(1-\alpha\right)$ for a certain sample, then the sample is an outlier, and vice versa.

2.3. The Principles and Algorithms of Rocke Estimator

Rocke estimator first proposed by Rocke is a robust estimation method for high-dimensional data based on an improved S-estimator, and then further improved and empirically compared to other estimates by Maronna and Yohai [17]. They pointed out that robustness was superior to other estimates when the data dimension was larger than 15. At the same time, the initial value for Rocke estimator has a significant influence. The subsampling approach usually employed for computing the starting values is very expensive for large dimensions.

This study demonstrates that a semi-deterministic equivariant procedure, initially proposed by Peñaand Prieto (2007) [18] for outlier detection, dramatically improves both the computing times and the statistical performances of the estimators. By comparing the initial values obtained by KSD estimator [18] with those obtained by MVE estimator, KSD estimatoris more robust and rapid than other initial values in terms of robustness and operation speed. For a set of samples $X_{n,p}$ with empirical distribution function F_n, we first estimate the initial mean vector μ and the covariance matrix Σ with KSD estimator, and calculate the squared of the Mahalanobis distance as $d_i = (x_i - \mu)^{\mathrm{T}} \Sigma^{-1} (x_i - \mu)$. Rocke estimator mainly applies a non-monotonic weight function by Rocke (1996) [19], and iteratively updates the weight of each sample point, and finally obtains a robust mean vector and covariance matrix estimator. When the distance change is very small, that is, the scale of the Mahalanobis distance to be small, the iteration is stopped, and a robust mean vector and covariance matrix can be obtained. The detailed algorithm for Rocke estimator is given by the following four steps:

In the first step, centering and scaling the data, and then the mean vector μ_0 and covariance matrix V_0 of the sample data are obtained by KSD estimator.

In the second step, let $\hat{\sigma} = \hat{\sigma}(d_1, d_2, \cdots, d_n)$ represent the scale estimate of the Mahalanobis distance, and solve it by

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{d_i}{\hat{\sigma}}\right) = \delta \tag{2}$$

where $\delta \in (0,1)$ controls the size of the breakdown point, when

$\delta = \frac{1}{n}\left[\frac{n-p+1}{2}\right]$, Rocke estimator can achieve the highest finite sample breakdown, where p represent the data dimension. The relationship between the ρ function and the weight function W is: $\rho' = W$, The function ρ is given by:

$$\rho = \begin{cases} 0, & 0 \le t \le 1-\gamma \\ \left(\frac{t-1}{4\gamma}\right)\left[3 - \left(\frac{t-1}{\gamma}\right)^2\right] + \frac{1}{2}, & 1-\gamma < t < 1+\gamma \\ 1, & t \ge 1+\gamma \end{cases} \tag{3}$$

where γ denotes the weight range. Since the Mahalanobis distance d approximates the chi-square χ_p^2 distribution with the degree of p, when the value of $\frac{d}{\hat{\sigma}}$ is outside $\left[\chi_p^2(\alpha), \chi_p^2(1-\alpha)\right]$, there is $W\left(\frac{d}{\hat{\sigma}}\right) = 0$ ($1-\alpha$ is the confidence). When the p is large, the χ_p^2 distribution tends to be symmetric, and there are:

$$\chi_p^2(0.5) \approx p \tag{4}$$

$$\chi_p^2(1-\alpha) - p \approx p - \chi_p^2(\alpha) \tag{5}$$

let $\gamma = \min\left(\frac{\chi_p^2(1-\alpha)}{p} - 1, 1\right)$, and calculate $\hat{\sigma}$ by fixed point method.

In the third step, the Mahalanobis distance is obtained by the initial value; then the new mean vector and the covariance matrix are calculated by the following weight function:

$$W(t)=\frac{3}{4\lambda}\left[1-\left(\frac{t-1}{\lambda}\right)^2\right]I(1-\gamma\le t\le 1+\gamma) \tag{6}$$

The different weights are applied to different samples by the following equation to calculate μ, the uncoordinated covariance matrix C to obtain the final covariance matrix Σ:

$$\sum_{i=1}^{n}W\left(\frac{d_i}{\hat{\sigma}}\right)(x_i-\mu)=0 \tag{7}$$

$$\frac{1}{n}\sum_{i=1}^{n}W\left(\frac{d_i}{\hat{\sigma}}\right)(x_i-\mu)(x_i-\mu)^{\mathrm{T}}=C \tag{8}$$

$$\Sigma=\frac{C}{|C|^{\frac{1}{p}}} \tag{9}$$

In the fourth step, repeat the second and third steps until obtain $\hat{\sigma}_{new}$ has the following relationship with the $\hat{\sigma}_{old}$ obtained last time: $\hat{\sigma}_{old}-\hat{\sigma}_{new}<tol$ stops the iteration, where tol is the preset error, and finally obtains the final stable mean vector and covariance matrix, and then calculates the Mahalanobis distance of the sample data.

3. Numerical Simulation Examples

We will apply the classical Mahalanobis distance and Mahalanobis distance based MCD estimator, and Mahalanobis distance based on the Rocke estimator to detect the outlier via the simulation studies. We generate mixture distribution data set, which consists of the standard normal distribution and contaminated data. The mixture distribution data is

$$N(0,1)+\varepsilon\left(\lambda N(0,1)+10_p\right)$$

where ε denotes the contaminated ratio and the constant λ determines the scatter of the outliers. In this simulation studies, we consider n = 100, p = 6; n = 300, p = 30, $\varepsilon=0,0.2$, and $\lambda=0,0.5$, and take $\chi_p^2(0.99)$ as the threshold. We calculate the above three Mahalanobis distance, and then identify the outliers in the data. We also calculate the number of real outliers (NRO), detected the corrected number of outliers (DENO), rate containing outliers in the data (DaOR), and detected outliers rate (DeOR). The corresponding results are shown in Figures 1-8.

It can be seen from the above results that under the 6-dimensional data setting, the results detected by the two robust Mahalanobis distances are substantially consistent with those of the classical Mahalanobis distances in the absence of outliers. When the outlier ratio increases to 20%, the classical Mahalanobis distances is affected by the outliers. Therefore, the classical Mahalanobis distance

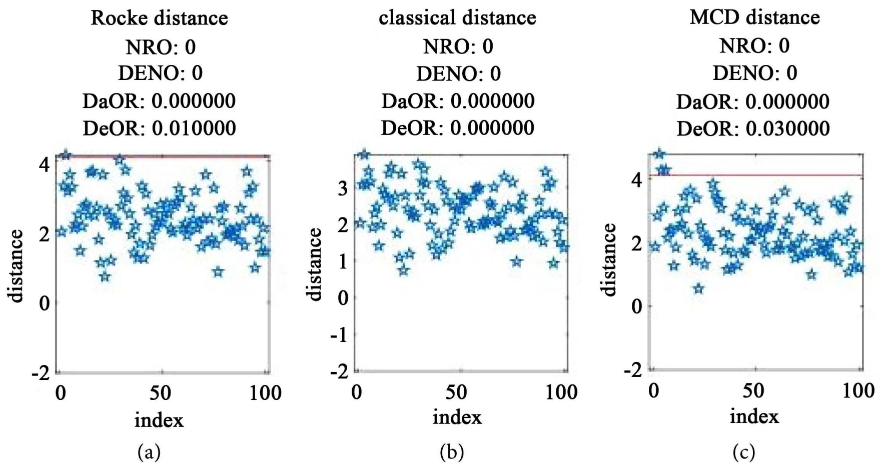


Figure 1. $\lambda = 0$, $n = 100$, $p = 6$, $\varepsilon = 0$. (a) Rocked detection; (b) Classical detection; (c) MCD detection.

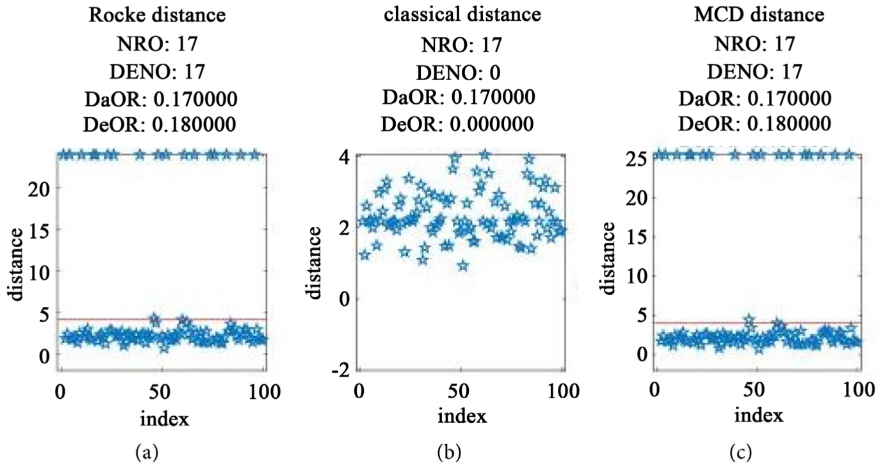


Figure 2. $\lambda = 0$, $n = 100$, $p = 6$, $\varepsilon = 0.2$. (a) Rocked detection; (b) Classical detection; (c) MCD detection.

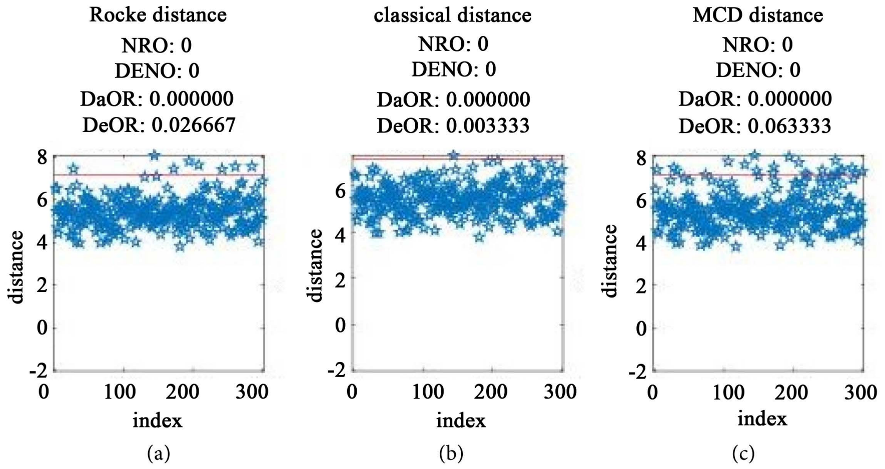
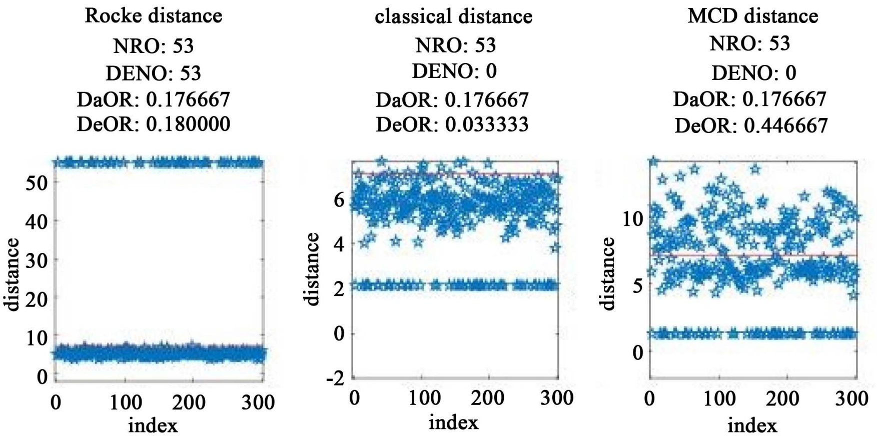
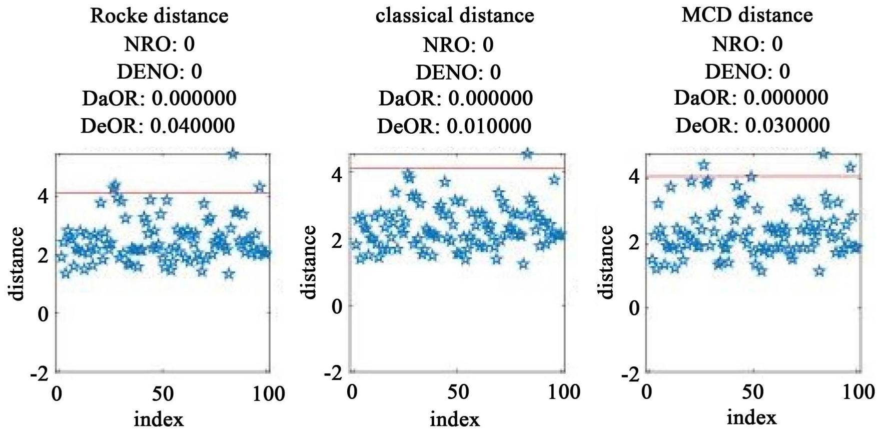


Figure 3. $\lambda = 0$, $n = 300$, $p = 30$, $\varepsilon = 0$. (a) Rocked detection; (b) Classical detection; (c) MCD detection.



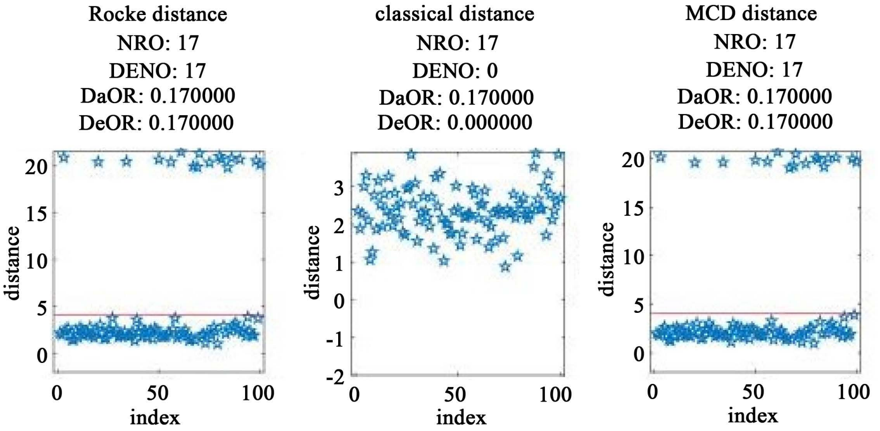
(a) (b) (c)

Figure 4. $\lambda = 0$, $n = 300$, $p = 30$, $\varepsilon = 0.2$. (a) Rocked detection; (b) Classical detection; (c) MCD detection.



(a) (b) (c)

Figure 5. $\lambda = 0.5$, $n = 100$, $p = 6$, $\varepsilon = 0$. (a) Rocked detection; (b) Classical detection; (c) MCD detection.



(a) (b) (c)

Figure 6. $\lambda = 0.5$, $n = 100$, $p = 6$, $\varepsilon = 0.2$. (a) Rocked detection; (b) Classical detection; (c) MCD detection.

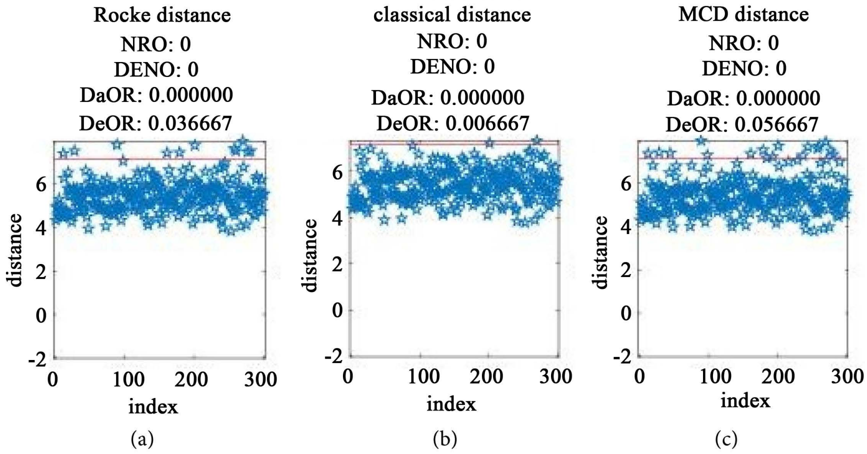


Figure 7. $\lambda = 0.5$, $n = 300$, $p = 30$, $\varepsilon = 0$. (a) Rocked detection; (b) Classical detection; (c) MCD detection.

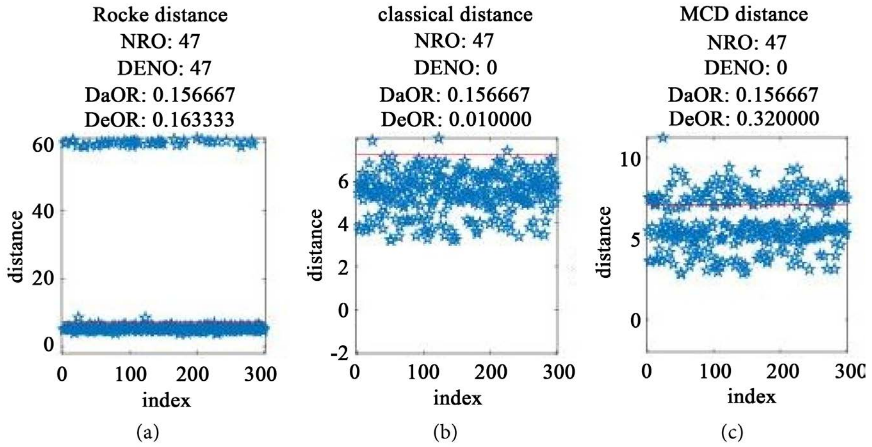
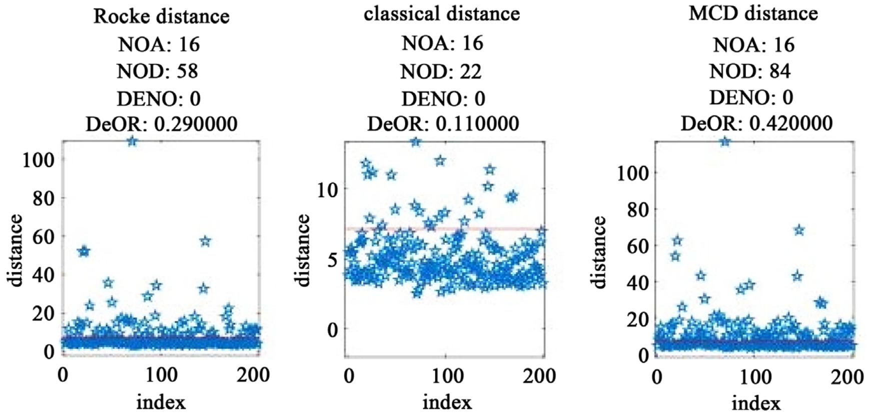


Figure 8. $\lambda = 0.5$, $n = 300$, $p = 30$, $\varepsilon = 0.2$. (a) Rocked detection; (b) Classical detection; (c) MCD detection.

cannot detect outliers. However, both robust Mahalanobis distances still maintain good robustness and can accurately detect outliers. In the 30-dimensional data setting, the detection results of the three Mahalanobis distances are also similar when there are no outliers. After the outlier ratio is increased to 20%, the classical Mahalanobis distance is still unable to detect the outliers. In addition, the detection result of the Mahalanobis distance based on MCD estimator has been greatly deviated. However, the Mahalanobis distance based on Rocke estimator can still accurately detect the outliers. Therefore, the Mahalanobis distance based on Rocke estimator can accurately detect the outliers in both low-dimensional data set and high-dimensional data set.

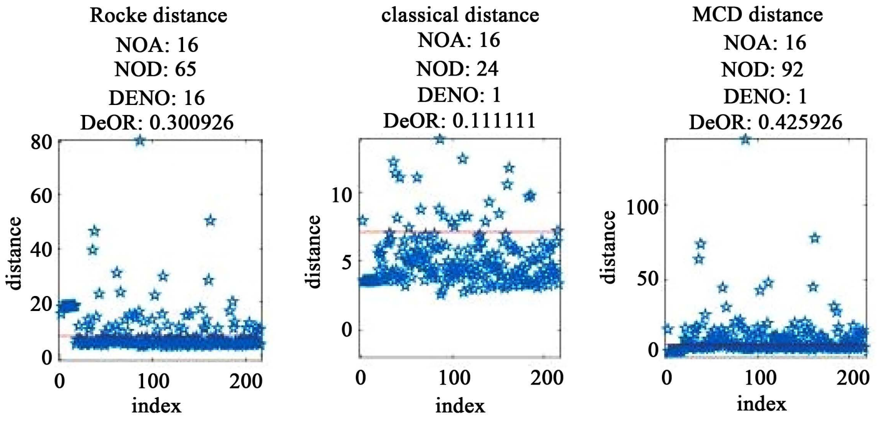
4. Empirical Analysis

In this section, we apply the proposed methodology to analyze the Breast Cancer Wisconsin (Diagnostic) Data Set (1995) [20]. The data set contained 30 test variables with a total of 569 data, of which 357 were diagnosed as benign and 212 were diagnosed as malignant. Classification by characteristic variables of the sample can distinguish between benign and malignant (Wolberg, Street, Heiseyand Mangasarian) [21]. Therefore, for data diagnosed as benign, data diagnosed as malignant is equivalent to the contaminated data. Adding contaminated data to the data diagnosed as benign with data diagnosed as malignant, and use $\chi_p^2(0.99)$ as the threshold. In the following, we apply the above three methods to detect outliers, and obtain the number and proportion of detected outlier and the scatter plot. Since the actual data may contain a certain proportion of outlier, it is not advisable to add too much data when the data diagnosed as malignant is used as the outliers. The first 200 diagnosed as benign data were taken, and 0 and 16 (100th to 115th) of the data diagnosed as malignant were sequentially added. We also calculate the number of outliers added (NOA), the number of outlier detected (NOD), detected the corrected number of outliers (DENO), and detected outliers rate (DeOR). The results are shown in **Figure 9** and **Figure 10**.



(a) (b) (c)

Figure 9. Diagnosed as benign data test results. (a) Rocked detection; (b) Classical detection; (c) MCD detection.



(a) (b) (c)

Figure 10. Adding diagnosed as malignancy data test results. (a) Rocked detection; (b) Classical detection; (c) MCD detection.

It can be observed that when no contaminated data is added, the proportion of outlier detected by Mahalanobis distance based MCD estimator, and Mahalanobis distance based on the Rocke estimator are about 30% and 40%, respectively, and the classical Mahalanobis distance detects about 10%. This implies that the data itself contains outliers. When 16 contaminated data are added, the classical Mahalanobis distance and Mahalanobis distance based MCD estimator can only accurately detect one of the real outliers added, but Mahalanobis distance based on the Rocke estimator can accurately detect the all 16 added outliers. Therefore, in the real data analysis, the Mahalanobis distance based on the Rocke estimator can be used to effectively identify the outliers in the high-dimensional data.

5. Discussion and Conclusion

Since the classical Mahalanobis distance is greatly affected by the outliers, there is a large deviation in either low-dimensional data set or high-dimensional data set when there are outliers in the data set. Therefore, the outliers cannot be accurately detected. Mahalanobis distance based MCD estimator can detect outliers more accurately in low-dimensional data, but it will produce large deviation in high-dimensional data. Our proposed method is more robust in both low-dimensional data set and high-dimensional data set. Specially, the robustness advantage is more obvious in high-dimensional data set, and it can accurately detect outliers. Through numerical simulation and empirical analysis, the accuracy and practicability of our proposed methodology in high-dimensional data set are validated. Thus, our proposed methodology can provide a new robust method for effectively detecting outliers in high-dimensional real data analysis.

Funding

Li's research is partially supported by and the Natural Science Foundation of Guangdong Province (No. 2018A030313171) and National Statistical Scientific Research Center Projects (No. 2017LY65).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Tukey, J.W. (1977) Exploratory Data Analysis (Vol. 2).

[2] Drineas, P., Mahoney, M.W. and Muthukrishnan, S. (2006) Sampling Algorithms for l 2 Regression and Applications. *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1127-1136. https://doi.org/10.1145/1109557.1109682

[3] Drineas, P., Magdon-Ismail, M., Mahoney, M.W. and Woodruff, D.P. (2012) Fast

Approximation of Matrix Coherence and Statistical Leverage. *Journal of Machine Learning Research*, **13**, 3475-3506.

[4] Drineas, P., Mahoney, M.W., Muthukrishnan, S. and Sarlós, T. (2011) Faster Least Squares Approximation. *Numerischemathematik*, **117**, 219-249. https://doi.org/10.1007/s00211-010-0331-6

[5] Mahalanobis, P.C. (1936) On the Generalized Distance in Statistics. *National Institute of Science of India*, **2**, 49-55.

[6] Holland, J.H. (1975) Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.

[7] Yan, Z., Dai, X.W. and Tian, M.Z. (2016) Diagnosis of Outlier Based on Sampling of Lever Large Dataset. *Mathematical Statistics and Management*, **35**, 794-802.

[8] Shi, D.D. Jia, R.Y. and Huang, Y.T. (2009) Improvement of Outlier Detection Algorithm in High Dimension Based on Genetic Algorithm. *Journal of Computer Technology and Development*, **19**, 141-143.

[9] Stahel, W.A. (1981) Robusteschätzungen: Infinitesimaleoptimalität und schätzungen von kovarianzmatrizen. Doctoral Dissertation, ETH, Zurich.

[10] Donoho, D.L. (1982) Breakdown Properties of Multivariate Location Estimators. Technical Report, Harvard University, Boston.

[11] Rousseeuw, P.J. (1985) Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications*, **8**, 37. https://doi.org/10.1007/978-94-009-5438-0_20

[12] Grübel, R. (1988) A Minimal Characterization of the Covariance Matrix. *Metrika*, **35**, 49-52. https://doi.org/10.1007/BF02613285

[13] Rousseeuw, P. and Yohai, V. (1984) Robust Regression by Means of S-Estimators. In: Franke, W.H.J. and Martin, D., Eds., *Robust and Nonlinear Time Series Analysis*, Springer, New York, 256-272. https://doi.org/10.1007/978-1-4615-7821-5_15

[14] Rousseeuw, P.J. and Driessen, K.V. (1999) A Fast Algorithm for the Minimum Covariance determinant Estimator. *Technometrics*, **41**, 212-223. https://doi.org/10.1080/00401706.1999.10485670

[15] Wang, B. and Chen, Y. (2005) Multivariate Anomaly Detection Based on Robust Mahalanobis Distance Based on. *Statistics and Decision*, 03X, 4-6.

[16] Feng, L., Li, B. and Huang, L. (2014) Detection and Analysis of Lidar Point Cloud Gross Error Based on Robust Mahalanobis Distance. *Geodesy and Geodynamics*, **34**, 168-173.

[17] Maronna, R.A. and Yohai, V.J. (2017) Robust and Efficient Estimation of Multivariate Scatter and Location. *Computational Statistics & Data Analysis*, **109**, 64-75. https://doi.org/10.1016/j.csda.2016.11.006

[18] Peña, D. and Prieto, F.J. (2007) Combining Random and Specific Directions for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data. *Journal of Computational and Graphical Statistics*, **16**, 228-254. https://doi.org/10.1198/106186007X181236

[19] Rocke, D.M. (1996) Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension. *The Annals of Statistics*, **24**, 1327-1345. https://doi.org/10.1214/aos/1032526972

[20] Wolberg, W.H., Street, W.N. and Mangasarian, O.L. (1992) Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/

[21] Wolberg, W.H., Street, W.N., Heisey, D.M. and Mangasarian, O.L. (1995) Computer-Derived Nuclear Features Distinguish Malignant from Benign Breast Cytology. *Human Pathology*, **26**, 792-796. https://doi.org/10.1016/0046-8177(95)90229-5