Pak. J. Statist.
2025 Vol. 41(1), 35-50

A NEW MODIFIED ROBUST MAHALANOBIS DISTANCE BASED ON MRCD TO DIAGNOSE HIGH LEVERAGE POINTS

Mohammed Al-Guraibawi[1], **Saif Hosam Raheem**[2§]
and **Bahr Kadhim Mohammed**[2]

[1] Al-Furat Al-Awsat Technical University
Diwaniyah Technical Institute, Iraq
Email: dw.moh2@atu.edu.iq

[2] College of Administration and Economics
University of Al-Qadisiyah, Iraq
Email: saif.hosam@qu.edu.iq
bahr.mahemmed@qu.edu.iq

[§] Corresponding author

ABSTRACT

Identifying outliers is a critical task in data analysis across various fields, financial, economics, healthcare and others. Outliers are indeed data points that differ significantly from the bulk of the data, it can have various implications depending on the context. Effective identification of outliers requires resistant statistical methods and a deep understanding of the context in which the data was generated. By accurately pinpointing outliers, analysts can make informed decisions, improve models, and gain deeper insights into the underlying processes driving the data. When the model contains multiple outliers and (or) high leverage points, the problem of masking and swamping arises. With this problem, the existing robust methods fail to identify outliers accurately and then make a big misdiagnosis.

In this article, we developed a new procedure with an efficient cut-off point to increase the correctly identification of high leverage points. The new procedure is depend on the Robust Mahalanobis Distance based on the "Minimum Regularized Covariance Determinant" with a new threshold term. To exam the performance of the developed method, simulation study in different scenarios are designed.

KEY WORDS

Outliers, High leverage points, Diagnostic, Masking, Swamping.

1. INTRODUCTION

Outliers play a pivotal role in data analysis across various fields by challenging the normative patterns observed within datasets, while often perceived as anomalies or errors, outliers can carry significant implications, potentially influencing the outcomes of statistical analyses and predictive models. Therefore, their timely detection and careful handling are critical steps in ensuring the reliability and validity of data-driven insights [Habshah, et al., (2009), Hadi (1992) and Saleem, Aslam and Shaukat (2021)].

The identification of outliers involves distinguish observations that significantly deviate from the bulk of points within a dataset (Hubert, Debruyne and Rousseeuw, 2018). This process is guided by specific definitions and detection techniques tailored to the characteristics of the data and the objectives of the analysis. Different disciplines and applications may employ distinct methods for outlier detection, reflecting the diverse contexts in which outliers can manifest and their potential impact on analytical outcomes. In the multivariate regression model, outliers may located in the explanatory variable and then defined as a high leverage points (HLP) (Habshah, et al., 2009).

In literature, many good approaches have been suggested to identify outliers and HLPs, such as the hat matrix, cooks distance and mahalanobis distance. Unfortunately, the classical methods are not robust due to the sensitivity to distribution assumptions. Where, the classical diagnostic methods often assume that the data follow a specific distribution, such as normality [Alguraibawi, Midi and Imon (2015) and Imon (2005)]. If these underlying assumptions are violated (e.g., they are highly skewed or exhibit heavy tails), the classical methods may completely destroyed. In addition, the thresholds used to isolate outliers or HLPs may not be appropriate or not enough efficient, leading to incorrect diagnoses

The one of significant problems that diagnostic methods suffer from is the masking and swamping impact. Masking occurs when outliers are hidden or suppressed by other data points, while swamping happens when some inlier observations are detected wrongly as outliers, skewing the results. The most diagnostic classical method that commonly used to detect outliers and HLP is the Mahalanobis Distance (MD) approach [Alguraibawi, Midi and Imon (2015) and Leroy and Rousseeuw (1987)]. The MD approach is an Euclidean distance between two points in multivariate space. When the normal assumptions are met, the MD is a powerful tool in multivariate analysis for measuring distances and identifying outliers and HLP based on the correlation structure of the data [Leroy and Rousseeuw (1987) and Rousseeuw and Yohai (1984)].

For multivariate regression model, let X be a $(n \times p)$ design matrix, where, p is a number of variables and n is a size of sample. The multivariate regression model in is given by,

$$y = X\beta + \varepsilon \tag{1}$$

where y be an $n \times 1$ vector of response variable, β be an $p \times 1$ vector of the unknown regression coefficients to be estimated and ε be an $n \times 1$ random vector assumed to be independently identically distributed normal with mean zero and constant variance.

Let $X_i = (x_{\mathrm{i}1,}, x_{i2}, \dots, x_{ip})^t$ be the ith vector of X, then the estimated of location $(\hat{\mu}_x)$ and scale $(\hat{\Sigma}_x)$ parameters of X are given as [Habshah, et al., (2009), Hadi (1992) and Mamun et al., (2012)];

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \; i = 1,2, \dots, n \tag{2}$$

$$\hat{\Sigma}_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^t (x_i - \hat{\mu}_x) \tag{3}$$

Then the classical MD for the ith case is given as follows

$$MD_x = d\left(x, \hat{\mu}, \hat{\Sigma}\right) = \left[(x - \hat{\mu}_x)^t \hat{\Sigma}^{-1} (x - \hat{\mu}_x)\right]^{\frac{1}{2}} \tag{4}$$

The MD_x can also be expressed depend on the hat values as

$$MD_x = \left[(n-1)\left(h_{ii} - \frac{1}{n}\right)\right]^{1/2} \tag{5}$$

where, $h_{ii} = x_i^t (X^T X)^{-1} x_i$

The h_{ii} is depend on the actual observations, so it sensitive to outliers points. Since the MD values depend on the h_{ii} as shown in Equation 5, it also is sensitive to outliers that can affect the estimation of location and scale parameters (Alguraibawi, Midi and Imon, 2015). The classical MD has another limitation, where the multivariate normal distribution is may not always be appropriate for real-world datasets. Many robust mahalanobis distance techniques are developed to overcome the drawbacks of classical MD, however, most of these techniques are suffered from swamping effect [Alguraibawi, Midi and Imon (2015) and Imon (2005)].

The threshold term of MD values is distributed as $\left(\sqrt{\boldsymbol{\chi}^2_{\boldsymbol{p},\ \mathbf{0.95}}}\right)$. MD value exceeds threshold term is considered as high leverage point (Rousseeuw and Yohai, 1984).

2. ROBUST MAHALANOBIS DISTANCE TECHNIQUES

Rousseeuw (1990) pointed that the classical MD suffer from the masking effect. In masking, the presence of one or a few outliers hides the presence of other outliers. Where the multiple outliers may be no have large value of MD (Rousseeuw and Leroy, 1987). This drawback is due to that the classical MD is depends on the traditional mean vector and variance covariance matrix those are not resistant for outliers. Rousseeuw suggested using robust estimators for location and scale parameters rather than classical mean and covariance for computing MD. The Robust MD (RMD) is defined as [Rousseeuw and Van Zomeren (1990) and Saleem, Aslam and Shaukat (2021)];

$$RMD_x = d(x, \hat{\mu}_{robust}, \hat{\Sigma}_{robust}) \tag{6}$$

where, $\hat{\mu}_{robust}$ and $\hat{\Sigma}_{robust}$ are the robust estimate of locations and scale coefficients, respectively. Many approaches are used to find robust location and scale coefficients such as, MCD, MVE, M-estimator, MM-estimator and newly the MRCD estimate [Ghapor et al., (2015), Rousseeuw and Yohai (1984) and Rousseeuw and Van Zomeren (1990)].

Leroy and Rousseeuw (1987) suggested using the term $\sqrt{\chi^2_{p,0.95}}$ as cut-off point for RMD, the using of this cut-off point is based on the assumption that the p-dimensional variables follow a multivariate normal distribution (Rousseeuw and Yohai, 1984). Since in real practice, there is no guarantee that this assumption be met, Imon and Khan (2003) suggested a cut-off point value as Imon (2005);

$$\text{median}(\text{RMD}_\text{i}) + \text{c}\ \text{mad}(\text{RMD}_\text{i}), i = 1,2,\ldots,n \tag{7}$$

Imon and Khan (2003) showed that the new cut-off point is more effective in diagnosing HLP than the cut-off point proposed by Leroy and Rousseeuw (1987).

3. ROBUST ESTIMATES OF LOCATION AND SCATTER PARAMETERS

The classical location and scale parameters have optimal properties under the normality assumptions. However, presence of small fraction of contamination in a sample data can make a big meaningless influence on the mean and variance of the sample due to these estimators are highly sensitive to outlying observations in a data. As an alternative method, robust statistical techniques are developed to be more resistant to unusual data. In literature, many robust estimates have been suggested such as MCD, MVE, M-estimator, MM-estimator OGK-estimator and recently, the MRCD [Ghapor et al., (2015), Imon and Khan (2003), Rousseeuw and Yohai (1984), Rousseeuw and Van Zomeren (1990) and Rousseeuw and Leroy (1987)].

In this study we will apply some of these robust estimator to robustify the MD to be more resistant to high leverage points.

3.1 The Minimum Covariance Determinant (MCD)

The "minimum covariance determinant" (MCD) proposed by Rousseeuw and Yohai (1984) is a robust estimator of a multivariate location and scale that has high breakdown point. The MCD aims to find h points (out of n) of dataset that gives the minimum determination of the scale matrix. The resampling technique is using to find h subset, where $\frac{n}{2} \leq h \leq n$. The robust location estimator $\hat{\mu}_x$ is determine as the average of h subset. Whereas, the robust scatter parameter $\hat{\Sigma}_x$ is the corresponding covariance matrix multiplied by a consistency factor such as C_α, where $\alpha = \frac{n-h}{h}$. The MCD estimator has high break down point around 0.50, but unfortunately, it has low efficiency at the normal model. Hubert et al., showed that the MCD estimator has only 6% relative efficiency with $\alpha = 0.50$ when $p = 2$ and 20.5% if $p = 10$. Another drawback for MCD estimator, the scatter matrix will be singular if $p > n$ leading to the determination value equal zero [Imon and Khan (2003), Rousseeuw and Yohai (1984) and Rousseeuw and Van Zomeren (1990)].

3.2 Minimum Volume Ellipsoid Estimator (MVE)

The "Minimum Volume Ellipsoid Estimator" (MVE) is the one of most popular robust estimator technique which suggested by Rousseeuw and Yohai (1984). The MVE estimator depend on finding the smallest ellipsoid that covers at least half points h (out of n) of observations. The algorithm of MVE has two steps. The first step is to identify the h subset that should be greater than half of data and the second step is to determine the minimum volume ellipsoid that covers at least the half of data. This technique lead to achieve a high break down point (around 50%), whereas, the efficiency of MVE is increasing by increase of h subset [Rousseeuw and Yohai (1984), Rousseeuw and Van Zomeren (1990) and Yohai (1987)].

3.3 M-Estimation

The M-estimators proposed by Huber (1964, 1973). It is a generalization of maximum-likelihood estimator aims to find estimators by reducing the effect of unusual data throw minimizing the sum of a less rapidly increasing function of residuals, as follows

$$\min_{\beta}\sum_{i=1}^{n}\gamma(r_i)=\min_{\beta}\sum_{i=1}^{n}\gamma\left(y_i-\sum_{i=1}^{n}x_{ij}\hat{\beta}_j\right) \tag{8}$$

where, r_i is the residual for $i=1,2,\dots,n$ and $\hat{\beta}$ is an unknown coefficients should be estimated. The objective γ is a particular function determines the contributions of each residuals in the objective function. The γ function must be positive definite, symmetric, unique minimum at zero and monotone in $|r_i|$. Under main assumptions, the M-estimates has about 95% relative efficiency with high breakdown points equal to 0.50. Although the M-estimator is robust for outlying observation in response variable but it is sensitive to high leverage points and the breakdown point will decreases if there is there is an outlier in the predictor variables [Alguraibawi, Midi and Imon (2015), Mamun et al., (2012) and Rousseeuw and Yohai (1984).

3.4 MM-Estimator

The MM-estimator is proposed by Yohai (1987). It is one of the most widely used robust estimation methods due to its many good features. It combines an elevated breakdown point 0.5 and supreme relative efficiency (95%). The Iterative reweighted least square (RWLS) approach is employed to obtain the MM-estimator. The "MM" computed throw using more than one M-estimation process to find the final estimates. The MM-estimates procedure is summarized as follows [Mamun et al., (2012) and Yohai (1987)];

1. Finding the initial estimates of the coefficients and the corresponding residuals $e_i, i=1,2,\dots,n$ depend on a high BP estimator such as S-estimators.
2. Computing the M-estimation of the scale of residuals $\hat{\sigma}_e$ using the results from Step 1.
3. By using the residuals e_i and the scale $\hat{\sigma}_e$ that obtained from previous steps and employed the first iteration of RWLS to find the M-estimates of the regression parameters based on Huber or bisquare weights φ_i

$$\sum_{i=1}^{n}\varphi_i\left(e_i^{(1)}/\hat{\sigma}_e\right)x_i=0 \tag{9}$$

4. Finding a new weights φ by using e_i from Step 3.
5. The $\hat{\sigma}_e$ is kept fixed from Step 2, Steps 3 and 4 are reiterated until convergence.

3.5 Minimum Regularized Covariance Determinant (MRCD)

We mentioned previously that the MCD estimators have essential restrictions that is of low efficiency and not available in high dimensional data when $p>n$. Boudt, K. (2020) proposed a new approach as an amendment for the MCD namely the "Minimum Regularized Covariance Determinant" (MRCD) estimator to avoid the MCD estimators drawbacks (Boudt et al., 2020). The main idea for MRCD is by replacing the subset based variance with a regularized variance estimate, which is specified as a weight average of the sample variance of the h subset and a pre-determined positive definite target matrix. The regularize variance based on the h subset, which results in the smallest overall determinant, is then the MRCD estimator (Imon and Khan, 2003). The minimum regularized covariance determinant is typically associated with robust covariance estimation, specifically in the context of robust statistics and outlier detection. The regularized covariance determinant can be formulated as [Boudt et al., (2020) and Zahariah and Midi (2023)]:

$$min_{\Sigma>0}|\Sigma|^{\frac{1}{n}} \tag{10}$$

where, $|\Sigma|$ represents the determinant of Σ, which is a measure of the volume or spread captured by the covariance matrix.

The objective $|\Sigma|^{1/n}$ is the regularized version of the covariance determinant. Regularization is often introduced to stabilize the estimation process, especially when dealing with high-dimensional data or when the sample size is relatively small compared to the number of variables

The MRCD estimator is formulated as [Boudt et al., (2020) and Zahariah and Midi (2023)]:

$$min_{\Sigma>0}\left[|\Sigma|^{1/n} + \lambda. tr(\Sigma)\right] \tag{11}$$

where:

- $\Sigma > 0$ denotes that Σ is positive definite, ensuring it is a valid covariance matrix.
- λ is a regularization parameter that balances the regularization term with the determinant term.
- $tr(\Sigma)$ is the trace of Σ, which is the sum of its diagonal elements (the sum of variances).

The objective function $|\Sigma|^{1/n} + \lambda. tr(\Sigma)$ aims to find a covariance matrix Σ that minimizes the regularized determinant $|\Sigma|^{1/n}$ while also penalizing the trace of Σ. This penalty term helps to control the complexity of the covariance matrix and can improve the stability and robustness of the estimator, particularly in the presence of outliers or when $p > n$. The MRCD estimator is typically computed using optimization techniques such as convex optimization methods [Boudt et al., (2020) and Zahariah and Midi (2023)].

4. A NEW "ROBUST MAHALANOBIS DISTANCE" BASED ON "MINIMUM REGULARIZED COVARIANCE DETERMINANT"

Boudt et al. (2020) proposed A RMD based on the bust covariance matrix of "Minimum Regularized Covariance Determinant". Then the robust location and scale of MRCD are used to calculate RMD values as outliers identification technique (Boudt et al., 2020). Boudt et al. suggested using $\left(\sqrt{\chi^2_{p,\ 0.99}}\right)$ as a cutoff point for RMD values to diagnose outliers (Hubert (2022). Siti Zahariah and Habshah Zahariah and Midi, 2023 pointed out that the Robust MD based on MRCD has high effective for the identification of high leverage point in high dimensional sparse dataset (Boudt et al., 2020). While the performance of the method deteriorates with increasing the number of predictor variables. In this study we suggested a cutoff point for RMD based on MRCD to overcome the drawback of Boudt et al. The new cutoff point depend on Imon's (2003) cutoff that mentioned in Equation (7) as follows (Imon 2005);

$$\text{median}(\text{RMD}_{\text{MRCD}}) + c\ \text{mad}(\text{RMD}_{MRCD}), i = 1,2, \dots, n \tag{12}$$

where; RMD_{MRCD} is a mahalanobis distance based on covariance of MRCD.

The proposed method are applying to identify outliers in both cases, for $n > p$ and for high dimensional data when $p > n$.

To assess the new suggested method, simulation study with different scenarios are designed and comparing it performance with some existing method.

5. MONTE CARLO SIMULATION STUDY

In order to verify the performance of the proposed method and to know its effectiveness in diagnosing the HLPs in the linear model, a Monte Carlo simulation study will be used. The suggested method namely MRMD based on MRCD is compared with some of existing diagnostic method such as RMD based on MCD, RMD based on MVE, RMD based on M- estimator, RMD based on MM- estimator and RMD based on MRCD.

A good diagnostic method is one that accurately diagnoses high leverage points with the least percentage of masking and swamping. The regular observations of variables of simulation study are generated as a normal distribution according the following formula in R language;

$$x_{ij} = rnorm\ (n, p), i = 1,2, \dots, n, j = 1,2, \dots, p\ \dots \tag{13}$$

where, n is a size of sample and p is a number of variables

In addition, different size of samples, number of variables and percentage of contamination (α) are considered.

In the first simulation study, we consider two variable with $n = 100$. To generating high leverage points in dataset, the 5% and 10% points of the regular data in both variables being replaced with relative large fixed values equal to 5. Tables 1 and 2 showed that all of robust diagnostic method are correctly identification of HLPs without any masking or swamping point except the RMD_MRCD that which tends to swamp some low leverage points as shown in Figures 1 and 2.

Table 1
The First Scenario of Simulation Study for Diagnose HLPs with $n = 100$, $p = 2$ and $\alpha = 5\%$.

Methods	Right Diagnosis Rate for HLPs	Number of Masking Points	Number of Swamping Points
RMD_MCD	100%	0	0
RMD_MVE	100%	0	0
RMD_Mest.	100%	0	0
RMD_MMest.	100%	0	0
RMD_MRCD	100%	0	1
MRMD_MRCD	100%	0	0

Table 2
The First Scenario of Simulation Study for Diagnose HLPs with $n = 100$, $p = 2$ and $\alpha = 10\%$

Methods	Right Diagnosis Rate for HLPs	Number of Masking Points	Number of Swamping Points
RMD_MCD	100%	0	0
RMD_MVE	100%	0	0
RMD_Mest.	100%	0	0
RMD_MMest.	100%	0	0
RMD_MRCD	100%	0	2
MRMD_MRCD	100%	0	0

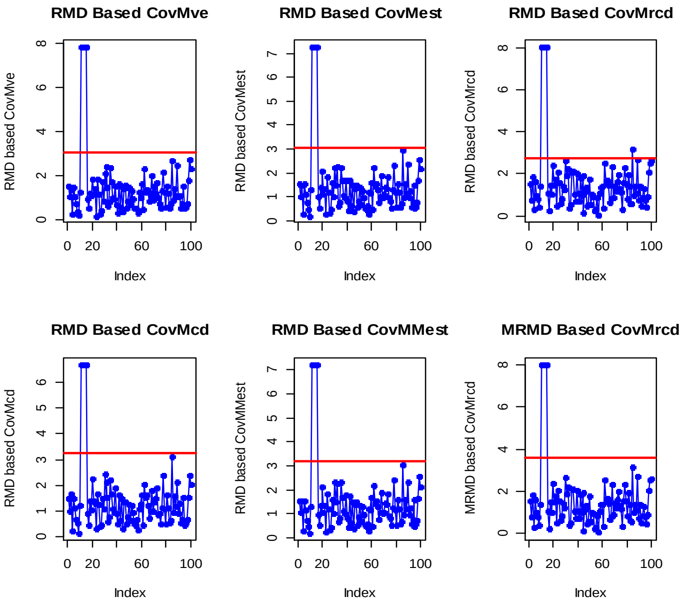

Figure 1: Plots of Robust Diagnostic Methods for Simulation Data with $n = 100$, $p = 2$ and $\alpha = 5\%$

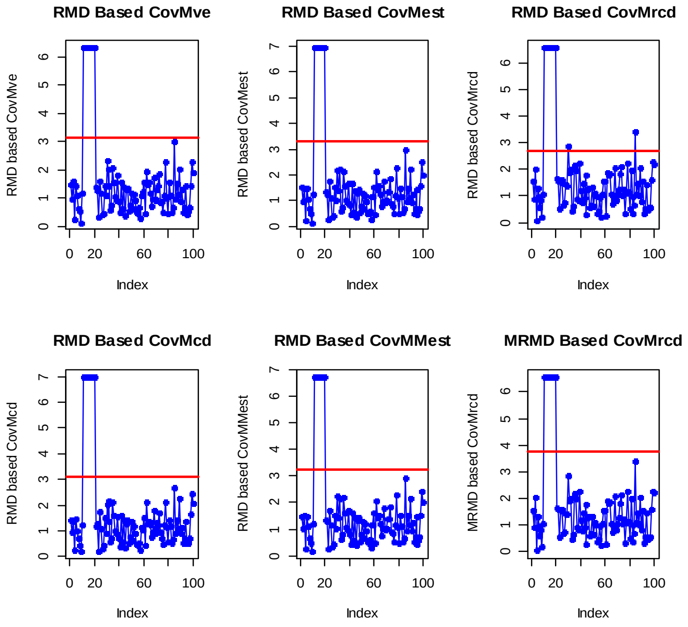

Figure 2: Plots of Robust Diagnostic Methods for Simulation Data with $n = 100$, $p = 2$ and $\alpha = 10\%$

In the second simulation study, we generate high dimension model with $n = 200$ and $p = 30$ with percentage of contamination $\alpha = 5\%$ and 10%. The high leverage points are generated by replacing the first α% regular values in the variables 1, 5 and 10 by large fixed value equal to 10. The results are summarized in Tables 3 and 4. It is clearly to see that the suggested method MRMD_MRCD has the supreme performance with 100% correctly identification followed by RMD_MMest which swamp just one points (point number 98) with $\alpha = 5\%$. Conversely, the RMD_MRCD has worst performance with 64 swamping points. Surprisingly, the RMD_MMest method completely collapse on high-dimensional data at 10% contamination. Figures 3 and 4 confirmed the obtained results.

Table 3
The Second Scenario of Simulation Study for Diagnose HLPs with $n = 200$, $p = 30$ and $\alpha = 5\%$

Methods	**Right Diagnosis Rate for HLPs**	**Number of Masking Points**	**Number of Swamping Points**
RMD_MCD	100%	0	8
RMD_MVE	100%	0	3
RMD_Mest.	100%	0	7
RMD_MMest.	100%	0	1
RMD_MRCD	100%	0	64
MRMD_MRCD	100%	0	0

Table 4
The Second Scenario of Simulation Study for Diagnose HLPs with $n = 200$, $p = 30$ and $\alpha = 10\%$

Methods	Right Diagnosis Rate for HLPs	Number of Masking Points	Number of Swamping Points
RMD_MCD	100%	0	6
RMD_MVE	100%	0	3
RMD_Mest.	100%	0	3
RMD_MMest.	10%	18	1
RMD_MRCD	100%	0	178
MRMD_MRCD	100%	0	0

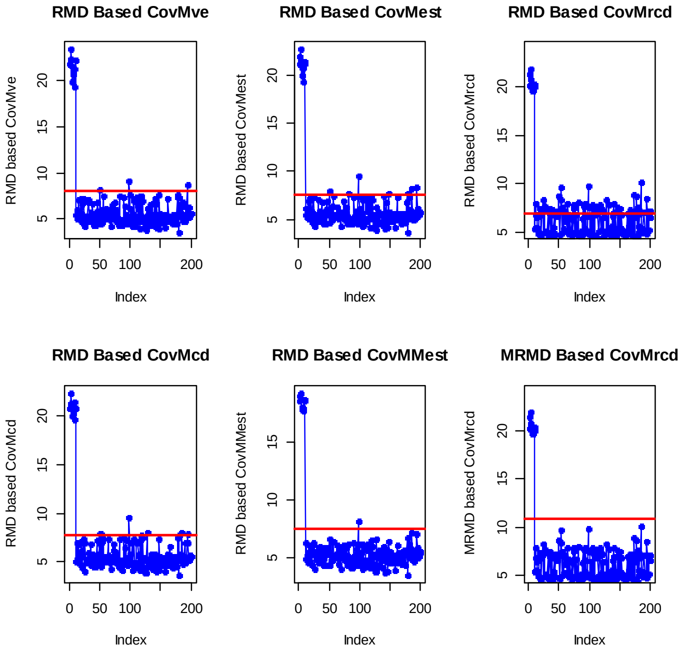

Figure 3: Plots of Robust Diagnostic Methods for Simulation Data with $n = 200$, $p = 30$ and $\alpha = 5\%$

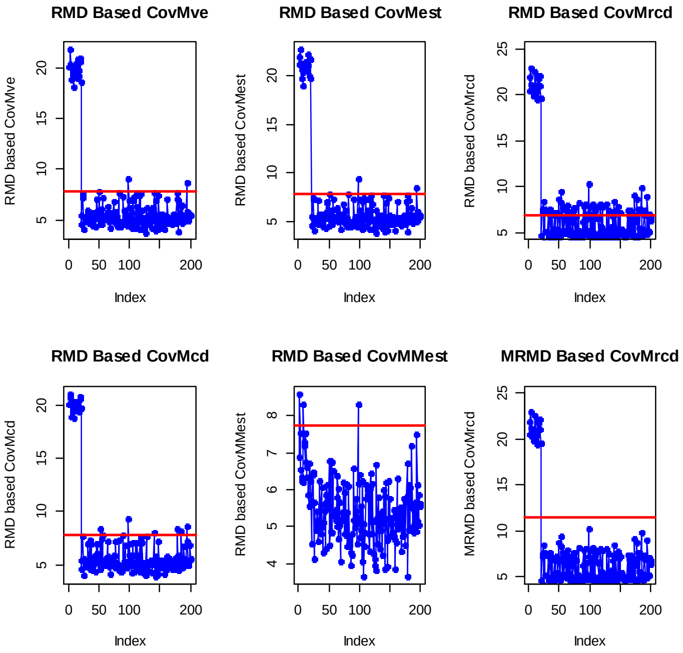

Figure 4: Plots of Robust Diagnostic Methods for Simulation Data with $n = 200$, $p = 30$ and $\alpha = 10\%$

In the third simulation study, we generate high dimension model with $n = 400$ and $p = 60$ with percentage of contamination $\alpha = 5\%$ and 10%. The high leverage points are generated by replacing the first α% regular values in the variables 1, 10, 20, 30 and 40 by large fixed value equal to 10. The results in Tables 5 and 6 and Figures 5 and 6 confirm that the proposed method MRMD_MRCD is still the best diagnostic method with stable performance with variations in the performance of the rest of the methods. We can also notice that the RMD_MMest method breakdown completely at a contamination of 10%.

Table 5
The Third Scenario of Simulation Study for Diagnose HLPs with $n = 400$, $p = 60$ and $\alpha = 5\%$

Methods	Right Diagnosis Rate for HLPs	Number of Masking Points	Number of Swamping Points
RMD_MCD	100%	0	19
RMD_MVE	100%	0	8
RMD_Mest.	100%	18	15
RMD_MMest.	100%	0	1
RMD_MRCD	100%	0	162
MRMD_MRCD	100%	0	0

Table 6
The Third Scenario of Simulation Study for Diagnose HLPs with $n = 400, p = 60$ and $\alpha = 10\%$

Methods	Right Diagnosis Rate for HLPs	Number of Masking Points	Number of Swamping Points
RMD_MCD	100%	0	17
RMD_MVE	100%	2	7
RMD_Mest.	100%	18	8
RMD_MMest.	2.5%	19	2
RMD_MRCD	100%	0	146
MRMD_MRCD	100%	0	0

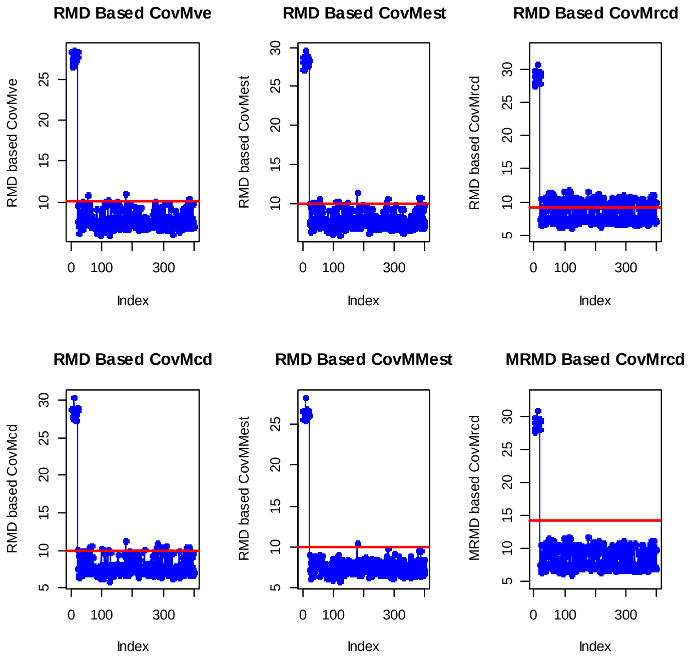

Figure 5: Plots of Robust Diagnostic Methods for Simulation Data with $n = 400, p = 60$ and $\alpha = 5\%$

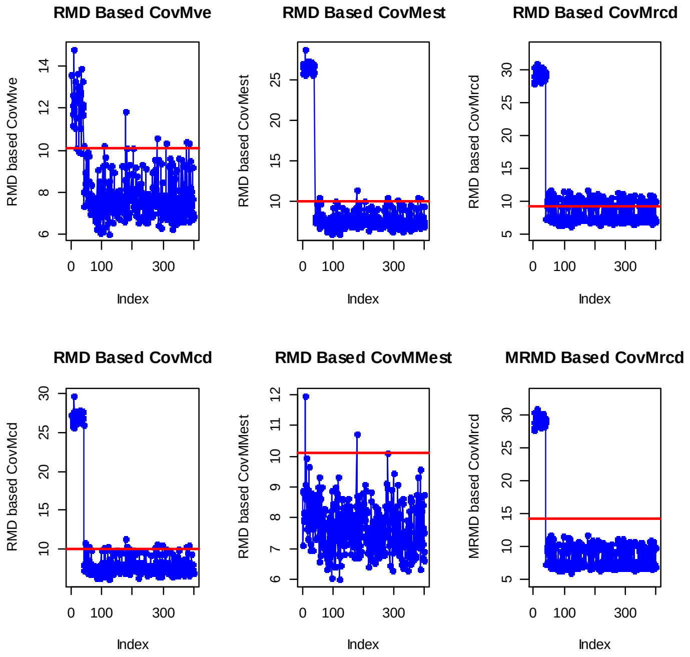

Figure 6: Plots of Robust Diagnostic Methods for Simulation Data with $n = 400$, $p = 60$ and $\alpha = 10\%$

In the last simulation study, we generate high dimension data when the number of variables are greater than the size of sample $(p > n)$ with $n = 200$ and $p = 400$. Two percentage of contamination $\alpha = 5\%$ and 10% are considered. The high leverage points are generated by replacing the first α% regular values in the first 10 variables with fixed value equal to 20. Its interested to show that all of existing method are breakdown when $(p > n)$, whereas MRMD_MRCD and MRMD_MRCD still working under this constriction. This issue is because that the MCD, MVE, M_estimator and MM_estimator are an efficient covariance estimator methods when the number of observations are smaller than number of variables, whereas, MRCD is efficient approach under this constriction. The results in Tables 7 and 8 and Figures 7 show that the proposed method MRMD_MRCD is the best diagnostic method with stable performance for all of contamination rates. Although the RMD_MRCD is identify all of HLPs, but it swamp a large number of regular points (more than 175 point).

Table 7
The Forth Scenario of Simulation Study for Diagnose HLPs with $n = 200$, $p = 400$ and $\alpha = 5\%$

Methods	Right Diagnosis Rate for HLPs	Number of Masking Points	Number of Swamping Points
RMD_MRCD	100%	0	190
MRMD_MRCD	100%	0	0

Table 8
The Forth Scenario of Simulation Study for Diagnose HLPs with $n = 200$, $p = 400$ and $\alpha = 10\%$

Methods	Right Diagnosis Rate for HLPs	Number of Masking Points	Number of Swamping Points
RMD_MRCD	100%	0	175
MRMD_MRCD	100%	0	0

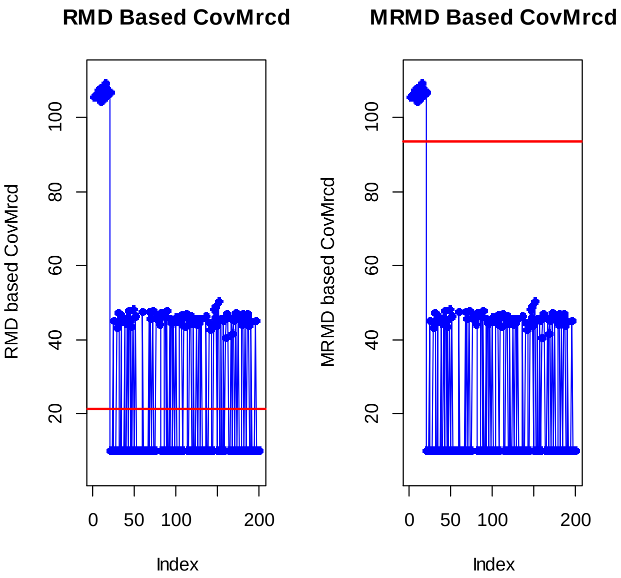

Figure 7: Plots of Robust Diagnostic Methods for Simulation Data with $n = 200$, $p = 400$ and $\alpha = 10\%$

7. CONCLUSIONS

The main target of this study is to suggest a new diagnostics method to identify high leverage points. The proposed method is a modification of robust mahalanobis distance based on MRCD with a new cut of point, named shortly MRMD_MRD. From the results of simulation study, we can conclude the following;

1. With low dimensions data, when $p = 2$ and different percentage of contaminations, we find that all of the robust diagnostic method are correctly identification of HLPs without any masking or swamping point except the RMD_MRCD that which tends to swamp some low leverage points.
2. With high dimensional data, when $p = 30$ and 60 with different sizes of samples, the proposed method has a supreme performance compared with existing methods that suffer from swamping problem

3. At 5% of contamination, the RMD_MRCD has the worst performance due to it has a high swamping points percentage.
4. At a contamination of 10%, the RMD_MMest method breakdown completely with unstable performance
5. When $(p > n)$, all diagnostic methods are destroyed, except the MRMD_MRCD and RMD_MRCD.

Finally, the proposed method of MRMD_MRCD has a perfect performance of diagnostic of high leverage points with reducing of masking and swamping effects for high dimensional data.

REFERENCES

1. Alguraibawi, M., Midi, H. and Imon, A.R. (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*, 2015(1), 279472.
2. Boudt, K., Rousseeuw, P.J., Vanduffel, S. and Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30(1), 113-128.
3. Ghapor, A., Zubairi, Y., Mamun, A.S.M.A. and Imon, A.H.M.R. (2015). A robust nonparametric slope estimation in linear functional relationship model. *Pak. J. Statist*, 31(3), 339-350.
4. Habshah, M., Norazan, M. and Rahmatullah Imon, A. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, 36(5), 507-520.
5. Hadi, A.S. (1992). Imon, A. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies*, 3, 207-218.
6. Hubert, M., Debruyne, M. and Rousseeuw, P.J. (2018). Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(3), e1421.
7. Hubert, M. (2022). Outlier detection in non-elliptical data by kernel MRCD. In *Statistics and Data Science Seminar, Location: Department of Mathematics and Statistics, Auburn University*.
8. Imon, A.H.M.R. and Khan, M.A.I. (2003). A solution to the problem of multcollinearity caused by the presence of multiple high leverage points. *International Journal of Statistical*, 2, 37-50.
9. Imon, A. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics*, 32(9), 929-946.
10. Leroy, A.M. and Rousseeuw, P.J. (1987). Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley*, 1987, 1
11. Mamun, A.S.M.A., Hussin, A.G., Zubairi, Y.Z. and Imon, A.H.M.R. (2012). A Nonparametric Robust Estimator for Slope of Linear Structural Relationship Model. *Pakistan Journal of Statistics*, 28(3), 385-394.
12. Rousseeuw, P.J. and Yohai, V. (1984). Robust regression by means of S-estimators, Robust and Nonlinear Time series Analysis. *Lecture Notes in Statistics*, 26, 256-272.
13. Rousseeuw, P.J. and Van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633-639.

14. Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
15. Saleem, S., Aslam, M. and Shaukat, M.R. (2021). A Review and Empirical Comparison of Univariate Outlier Detection Methods. *Pak. J. Statist*., 37(4), 447-462.
16. Yohai, V.J. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15, 642-656.
17. Zahariah, S. and Midi, H. (2023). Minimum regularized covariance determinant and principal component analysis-based method for the identification of high leverage points in high dimensional sparse data. *Journal of Applied Statistics*, 50(13), 28[3]17-2835.
18. AL-Sabbah, S.A., Mohammed, L.A. and Raheem, S.H. (2021). Sliced Inverse Regression (SIR) with Robust Group Lasso. *International Journal of Agricultural & Statistical Sciences*, 17(1), p359.