Zero: Jurnal Sains, Matematika, dan Terapan
E-ISSN: 2580-5754
P-ISSN: 2580-569X
Vol. 8, No. 1, June 2024
DOI: 10.30829/zero.v8i1.20808
pp. 40-46

Minimum Volume Ellipsoid (MVE) and Minimum Determinant Covariance (MCD) Methods for Estimating Covariance Matrix in Multivariate Data

Ananda Ifrajiani Khair[1], Sutarman[2]
[1,2]Department of Mathematics, Universitas Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received 05 27, 2024
Revised 06 27, 2024
Accepted 06 30, 2024

Keywords:

Covariance Matrix
MCD Method
Multivariate Data
MVE Method
Outlier

ABSTRACT

Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) are robust methods used to handle the outlier problem. Outliers are points that appear to deviate significantly from other data sample points that can have a significant effect on the results of the analysis, so a robust method is needed to solve this problem. MVE and MCD have a high breakdown point or level of resistance to outliers, which is 50%, so that it can overcome the influence of extreme outliers. Based on this research, it is known that by using the same data, the MVE and MCD methods produce more robust estimates that were not affected by outliers. The non robust method just found 10 outliers, while the MVE method found that were 276 data points detected as outliers and for the MCD method, the estimation result is with 257 data points detected as outliers.

Corresponding Author:

Ananda Ifrajiani Khair,
Department of Mathematics,
Universitas Sumatera Utara, Medan, Indonesia
Email: nandaifra189@gmail.com

1. INTRODUCTION

The covariance matrix is crucial in multivariate data analysis. It evaluates the dispersion of each variable and the degree of correlation between pairs of variables within the multivariate data set. [1]. Estimating the covariance matrix can pose a number of difficulties due to data of large size and dimension and the presence of outliers in the data. Outliers are points that appear to deviate markedly from other sample members in the data [2]. Outliers will be difficult to detect if non-robust methods are used. Therefore, robust methods are needed to reduce the impact of outliers on the result of statistical analysis.

Basically, previous studies on robust methods such as Least Trimmed Square (LTS), Least Median of Square (LMS), and Tukey M [3] and also LASSO regression [4] have shown that when analyzing multivariate data, robust methods are needed to handle outliers. However, robust methods have different breakdown points. A high breakdown point will be more able to tolerate extreme values or outliers, so as to provide consistent and reliable results when analyzing multivariate data[5]. Therefore, this study uses Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) methods on the covariance matrix because they have a high breakdown point. MVE is used to find the minimum ellipsoid that can express most of the data. By finding the minimum ellipsoid, MVE is able to identify important structures in the data, offering a more comprehensive understanding of the data's distribution and characteristics. Meanwhile, MCD is a method that identifies the subset of multivariate data that has the smallest covariance determinant value[5].

2. RESEARCH METHOD

2.1 Robust Method

Robust methods are a regression analysis approach employed when the residual distribution deviates from the normal pattern. These methods involve initially fitting the regression model to a subset of the data and subsequently managing the outliers, ensuring that the analysis remains unaffected by them. Thus, robust methods do not eliminate any data but instead identify an appropriate model from a portion of the data. [6]. The use of estimation in robust methods can vary in different situations, so it is important to choose the right estimation when analysing data. One way to do this is by considering the breakdown point (ε), which is the smallest fraction of the sample (compared to n) that can render the estimation useless [7].

2.2 Minimum Volume Ellipsoid (MVE)

Minimum Volume Ellipsoid (MVE) is the first robust method with a high breakdown point used as a method to detect outliers in multivariate data. MVE aims to find robust estimates of the covariance matrix and its location vector [8]. The algorithm of MVE [9] is carried out by:

a. Construct a subsample containing $p + 1$ observations indexed by $J = \{i_1, \dots, i_{p+1}\} \subset \{1, \dots, n\}$.
b. Calculate the subsample mean with the equation $T_J = \frac{1}{p+1} \sum_{j=1}^{p+1} x_{ij}$ and subsample covariance matrix with $S_J = \frac{1}{p} \sum_{j=1}^{p+1} \left(x_{ij} - T_J\right)\left(x_{ij} - T_J\right)^T$
c. Calculate the squared distance value D_J^2 with the $D_J^2 = \left[\left(x_i - T_J\right)^T \left(S_J\right)^{-1} \left(x_i - T_J\right)\right]_{h:n}$
d. Calculate the volume value of V_J with the equation $V_J = \left(\frac{D_J}{c}\right)^P \det\left(S_J\right)^{\frac{1}{2}}$ for generates an ellipsoid
e. Reapeat steps a-d for each J subsample, then select V_J with the smallest value
f. Finding the T_{MVE} and S_{MVE} value of the subsample with the smallest volume.

2.3 Minimum Covariance Determinant (MCD)

Minimum Covariance Determinant (MCD) is a robust method used as an alternative to the MVE method which also has a high breakdown point value. MCD is similar to MVE and has a same objective function. The difference between these two methods lies in there constraints, that MCD uses only h points for estimations, rather than an ellipsoid ecompassing h points [10]. The algorithm of MCD [5] is carried out by:

a. Taking h number of different observations to estimate the covariance matrix
b. Calculate the average of each subsample with the equation $T = \frac{1}{h} \sum_{i=1}^{h} w_i$ and subsample covariance matrix with $S = \frac{1}{h} \sum_{i=1}^{h} (x_i - T)(x_i - T)^t$
c. Calculate the relative distance value with $d(i) = \sqrt{(x_i - T)^t S^{-1} (x_i - T)}$ which is measured based on the Mahalanobis distance between the vector x_i and the center point T in the coordinates determined by the covariance matrix S.
d. Sort the relative distance values from the smallest to largest. If the determinant of the covariance matrix converges, then the iterations stops
e. Selecting the subsample with the smallest covariance matrix determinant
f. Find the value of T_{MCD} and S_{MCD} from the subsample with the smallest determinant.

2.4 Outliers

Outliers are points that appear to deviate markedly form other sample members in the data [2]. Another definition states that an outlier is an observation (or a subset of observations) that appears inconsistent with the rest of the data set [11]. Additionally, outliers are data points that are significantly different from other data points or do not conform to expected normal behavior (based on defined abnormal behavior) [12]. Outliers can also be patterns in the data that do not align with a clear concept of normal behavior [13]. Detecting outliers can be challenging because they can adjust themselves to appear normal, a phenomenon known as the masking effect. Furthermore, the swamping effect complicates outlier detection by the overwhelming presence of non-outlier data dominating the dataset. Despite outliers having significant values, their impact may appear relatively small compared to the majority of data, leading to errors in identifying outliers as normal [14]

2.5 Covariance Matrix

A covariance matrix is used to quantify the degree of covariance between two or more random variables. Suppose we have a dataset comprising n observations of two variables, X and Y:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \qquad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Then the data vector can be expressed as:

$$Z = [X \quad Y] = \begin{bmatrix} X_1 & Y_1 \\ X_2 & Y_2 \\ \vdots & \vdots \\ X_n & Y_n \end{bmatrix}$$

and the covariance matrix of Z is:

$$\Sigma = \begin{bmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{bmatrix}$$

with $cov(X,X)$ is the variance of the variable X, $cov(Y,Y)$ is the variance of the variable Y, and $cov(X,Y)$ is the variance of the variable X and Y [15].

3. RESULT AND ANALYSIS

3.1 Data Source

The data used is a concrete compressive strength dataset obtained from the Kaggle.com website. The dataset consists of $n = 1030$ observations and $p = 6$ variables, specifically on variables water, superplasticizer, coarse aggregate, fine aggregate, and age (day). Descriptive statistical calculations for the mean and covariance matrix were conducted using Python, yielding the following results:

- Mean:

$$\bar{X} = [281{,}165 \quad 181{,}566 \quad 6{,}203 \quad 972{,}918 \quad 773{,}578 \quad 45{,}662]$$

- Covariance Matrix:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

$$= \begin{bmatrix} 10921{,}742 & -181{,}989 & 57{,}914 & -888{,}608 & -186{,}151 & 540{,}991 \\ -181{,}989 & 456{,}060 & -83{,}870 & -302{,}724 & -771{,}573 & 374{,}496 \\ 57{,}914 & -83{,}870 & 35{,}682 & -123{,}687 & 106{,}562 & -72{,}720 \\ -888{,}608 & -302{,}724 & -123{,}687 & 6045{,}656 & -1112{,}795 & -14{,}811 \\ -1866{,}151 & -771{,}573 & 106{,}562 & -1112{,}795 & 6428{,}099 & -790{,}565 \\ 540{,}991 & 374{,}496 & -72{,}720 & -14{,}811 & -790{,}565 & 3990{,}437 \end{bmatrix}$$

3.2 Estimating Covariance Matrix with The MVE Method

The MVE method is used to estimate the covariance matrix by finding the smallest ellipsoid that contains most of the data points. This ellipsoid is a representation of the overall data pattern and is used to identify points that are considered outliers. In calculations carried out with the help of Python programming, the results yielded T_{MVE} and S_{MVE} for the subsample with the smallest volume, specifically:

- $T_{MVE} = [245{,}09 \quad 180{,}28 \quad 7{,}69 \quad 966{,}09 \quad 796{,}03 \quad 18{,}85]$
- $S_{MVE} = \begin{bmatrix} 4948{,}7 & -28{,}6 & 26{,}7 & -982{,}1 & -1409{,}6 & -273{,}9 \\ -28{,}6 & 328{,}09 & -42{,}4 & -591{,}7 & -449{,}5 & 58{,}2 \\ 26{,}7 & -42{,}4 & 29{,}3 & -188{,}8 & -35{,}2 & 8{,}4 \\ -982{,}1 & -591{,}7 & -188{,}8 & 6518{,}9 & -423{,}4 & -186{,}4 \\ -1409{,}6 & -449{,}5 & -35{,}2 & -423{,}4 & 5487{,}8 & -203{,}3 \\ -273{,}9 & 58{,}2 & 8{,}4 & -186{,}4 & -203{,}3 & 143{,}4 \end{bmatrix}$
- $S_{MVE}^{-1} = \begin{bmatrix} 0{,}0004 & 0{,}0016 & 0{,}0047 & 0{,}0003 & 0{,}0003 & 0{,}0007 \\ 0{,}0016 & 0{,}0259 & 0{,}0716 & 0{,}0048 & 0{,}0033 & -0{,}0004 \\ 0{,}0047 & 0{,}0716 & 0{,}2401 & 0{,}0147 & 0{,}0097 & -0{,}0011 \\ 0{,}0003 & 0{,}0048 & 0{,}0147 & 0{,}0011 & 0{,}0006 & 0{,}0003 \\ 0{,}0003 & 0{,}0033 & 0{,}0097 & 0{,}0006 & 0{,}0006 & 0{,}0005 \\ 0{,}0007 & -0{,}0004 & -0{,}0011 & 0{,}0003 & 0{,}0005 & 0{,}0099 \end{bmatrix}$

3.3 Estimating Covariance Matrix with The MCD Method

In contrast to the MVE method, the MCD method estimates using only h points rather than an ellipsoid encompassing h points. The results of iterations conducted using Python programming are as follows:

Table 3.1 Iterations of the MCD Method

Iterations	Determinant of Covariance Matrix
1	4,774144064340823e+18
2	2.7198523535066668e+16
3	1.6163399171017386e+16
4	1.1075033302575084e+16
5	1.0762317114815812e+16
6	1.0507447233827772e+16
7	1.0336751528594116e+16
8	1.0296755038890786e+16
9	1.0275624739294708e+16
10	1.0267275820726456e+16
11	1.0267275820726456e+16

After the iteration is complete, the subsample that has the smallest covariance matrix determinat value is selected. Table 3.1 show that the subsample with the smallest determinant value is at iteration 11. Then, from that subsample, the value is obtained:

- $T_{MCD} = [255{,}43 \quad 178{,}99 \quad 5{,}71 \quad 993{,}44 \quad 787{,}401 \quad 20{,}66]$
- $S_{MCD} = \begin{bmatrix} 5{,}5 & -1{,}4 & -3{,}3 & -8{,}8 & -1{,}3 & -1{,}1 \\ -1{,}4 & 2{,}1 & -5{,}3 & -1{,}6 & -1{,}6 & -2{,}4 \\ -3{,}3 & -5{,}3 & 2{,}2 & -7{,}01 & 3{,}9 & 1{,}6 \\ -8{,}8 & -1{,}6 & -7{,}01 & 4{,}4 & -3{,}2 & -3{,}9 \\ -1{,}3 & -1{,}6 & 3{,}9 & -3{,}2 & 2{,}3 & -7{,}9 \\ -1{,}1 & -2{,}4 & 1{,}6 & -3{,}9 & -7{,}9 & 2{,}05 \end{bmatrix}$
- ${S_{MCD}}^{-1} = \begin{bmatrix} 2{,}1 & 6{,}3 & 2{,}02 & 1{,}01 & 3{,}8 & 6{,}4 \\ 6{,}3 & 2{,}09 & 5{,}5 & 1{,}8 & 8{,}04 & -8{,}5 \\ 2{,}02 & 5{,}5 & 1{,}9 & 5{,}6 & 1{,}2 & -6{,}3 \\ 1{,}01 & 1{,}8 & 5{,}6 & 4{,}08 & 9{,}5 & -5{,}4 \\ 3{,}8 & 8{,}04 & 1{,}2 & 9{,}5 & 4{,}8 & 2{,}2 \\ 6{,}4 & -8{,}5 & -6{,}3 & -5{,}4 & 2{,}2 & 5{,}3 \end{bmatrix}$

3.4 Outlier Detection

To detect outliers, the squared distance value (RD) will be calculated for each observation with the formula:

$$RD = \sqrt{(x_i - T)^T S^{-1} (x_i - T)}$$

Subsequent, the threshold chi-square value (c) will be calculated with:

$$c = \sqrt{x^2_{k;1-\alpha}}$$

This c value is used to determine wheter an observation is considered an outliers or not. In this study, the degree of freedom $(k) = 6$ and significance level $(\alpha) = 0{,}5$. Therefore, the threshold chi-square value used is:

$$c = \sqrt{x^2_{6;0{,}5}} = 5{,}348$$

Then, the observation with a value of $RD > c$ are identified as outliers.

Using a Python program as a tool, the results of outlier detection are as follows:

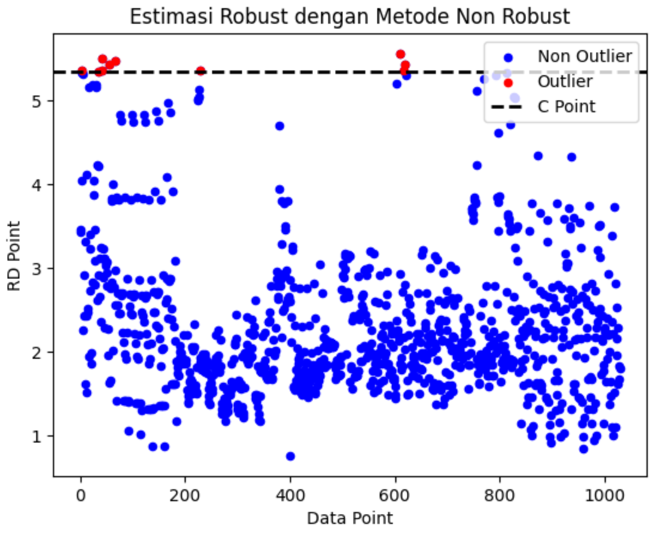

Figure 3.1 Non Robust Estimation

In Figure 3.1, the detection results using the non robust method show that there are only 10 outliers in the dataset. This shows that the non robust method is not able to overcome the influence of outliers, so that the estimations results do not show the true value.

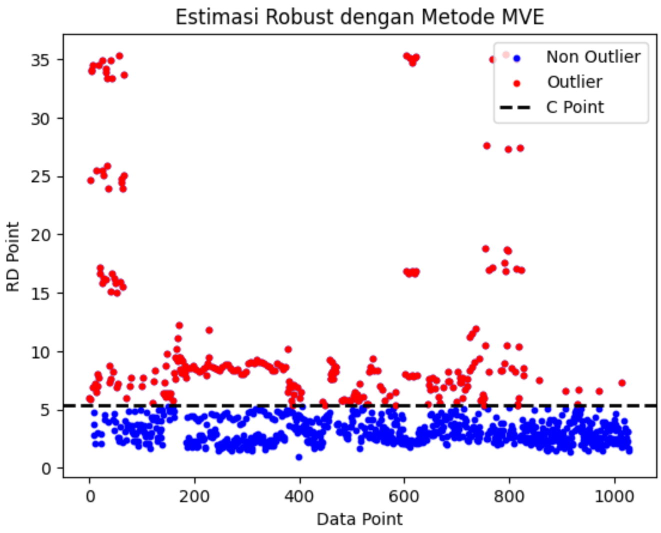

Figure 3.2 Robust Estimations with MVE Method

In Figure 3.2, it can be seen that the MVE method has successfully estimated the covariance matrix, so that the results are not affected by outliers. The outliers points depicted in red are points that are above the line of c point. In the calculations using Python, 276 data points were identified as outliers, while 754 data points were classified as non outliers.

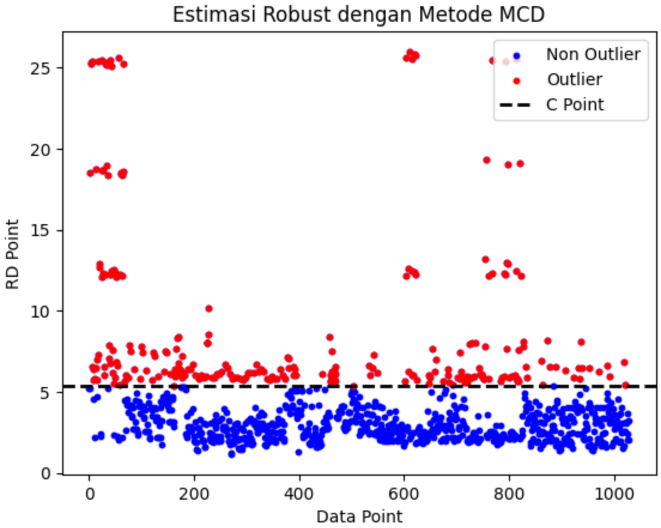

Figure 3.3 Robust Estimations with MCD Method

In Figure 3.3, it is evident that the MCD method has successfully estimated the covariance matrix, resulting in estimates that are not influenced by outliers, although the figure shows that the boundary between non-outlier points and outliers is very close. In the calculation using the Python application, 257 data points were identified as outliers, while 773 data points were classified as non-outliers.

4. CONCLUSION

Based on the discussion in the previous chapter, the results show that using the conventional method (non-robust method) detected only 10 outlier data points. In contrast, using the Minimum Volume Ellipsoid (MVE) method detected 276 outlier data points, while the Minimum Covariance Determinant (MCD) method detected 257 outlier data points. The success in detecting these outliers helps reduce the influence of extreme outliers, thereby mitigating masking and swamping effects and resulting in more accurate and robust estimates. Therefore, the findings from this research analysis indicate that the MVE and MCD methods perform well in estimating the covariance matrix for multivariate data.

REFERENCE

[1] G. M. Oyeyemi and R. A. Ipinyomi, “A robust method of estimating covariance matrix in multivariate data analysis,” *Analele Ştiinţifice ale Univ. »Alexandru Ioan Cuza« din Iaşi. Ştiinţe Econ.*, vol. 56, no. 1, pp. 586–601, 2009.

[2] F. E. Grubbs, “Procedures for Detecting Outlying Observations in Samples,” *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969, doi: 10.1080/00401706.1969.10490657.

[3] T. Zaman and H. Bulut, “Modified regression estimators using robust regression methods and covariance matrices in stratified random sampling,” *Commun. Stat. - Theory Methods*, vol. 49, no. 14, pp. 3407–3420, 2020, doi: 10.1080/03610926.2019.1588324.

[4] P. Pfeiffer and P. Filzmoser, “Robust statistical methods for high-dimensional data, with applications in tribology,” *Anal. Chim. Acta*, vol. 1279, no. August, p. 341762, 2023, doi: 10.1016/j.aca.2023.341762.

[5] Rousseeuw, P. J, and K. van Driessen, “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[6] F. P. Hidayatulloh, D. Yuniarti, and S. Wahyuningsih, “Regresi Robust Dengan Metode Estimasi-S,” *Eksponensial*, vol. 6, no. 2, pp. 163–170, 2015.

[7] S. F. Møller, J. Von Frese, and R. Bro, “Robust methods for multivariate data analysis,” *J. Chemom.*, vol. 19, no. 10, pp. 549–563, 2005, doi: 10.1002/cem.962.

[8] A. S. Hadi, “Identifying Multiple Outliers in Multivariate Data,” *J. R. Stat. Soc.*, vol. 54, no. 3, pp. 761–771, 1992.

[9] S. Van Aelst and P. Rousseeuw, “Minimum volume ellipsoid,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 1, no. 1, pp. 71–82, 2009, doi: 10.1002/wics.19.

[10] A. S. Hadi, A. H. M. Rahmatullah Imon, and M. Werner, “Detection of outliers,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 1, no. 1, pp. 57–70, 2009, doi: 10.1002/wics.6.

[11] V. Barnett and T. Lewis, “Outliers in statistical data, second edition,” *John Wiley Sons*, p. 463, 1994.

[12] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection : A Survey,” vol. 41, no. 3, pp. 1–58, 2009, doi: 10.1145/1541880.1541882.

[13] K. Singh and M. Cantt, “Outlier Detection : Applications And Techniques,” vol. 9, no. 1, pp. 307–323, 2012.

[14] A. Boukerche, L. Zheng, and O. Alfandi, “Outlier Detection: Methods, Models, and Classification,” *ACM Comput. Surv.*, vol. 53, no. 3, 2020, doi: 10.1145/3381028.

[15] G. Strang, *Introduction to Linear Algebra, Fourth Edition*, 4th Editio. Wellesley - Cambridge Press, 2009.