Detection of Outliers in Multivariate Data using Minimum Vector Variance Method

E T Herdiani¹, P P Sari¹ and N Sunusi¹

¹Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University
ntitisanusi@gmail.com

Abstract. Outliers are observations that do not follow the distribution of data patterns and can cause deviations from data analysis, so a method for identifying outliers is needed. One method in scanning detection is Minimum Vector Variance which is a robust estimator that uses the minimum Vector Variance (VV) criteria. In this study, the MVV method was used to detect outliers in criminality data in Indonesia in 2013 and data that had been entered out by 5% and 10%. The results showed that the MVV method was more effective than the Mahalanobis distance when detecting outliers in data that had been entered out by 5% and 10%.

1. Introduction
Detection of outliers especially in multivariate data is not easy. This is because problems become more complex when there are two or more outliers that come from more than two variables [1]. Outliers are observations that do not follow most patterns and are located far from the data center. The existence of outliers in the data can lead to deviations from the results of data analysis such as deviations from the results of statistical tests based on the average and covariance parameters. Therefore, it is necessary to identify its existence [2].

One method that is often used to detect outliers is the Mahalanobis distance, by calculating the distance of each observation of the data center. However, the Mahalanobis distance is still included in the classic estimator which still depends on basic assumptions such as normality, linearity, etc. [3]. Therefore, the hashtag method is needed for outliers. According to [4], one method that can be used in detecting outliers in multivariate data is Minimum Vector Variance (MVV). This method uses the minimum Vector Variance criteria to detect outliers. The reason for using MVV is because this method is more effective with a lower level of complexity and robustness to outliers [5].

2. Materials and Methods
2.1. Data Collection
Outliers identification in this study used multivariate data originating from the 2014 Republic of Indonesia Central Bureau of Statistics book on criminality data in Indonesia in 2013. Criminal data consisted of variables calculated from the number of crime cases that occurred per 100,000 population consisting of 31 observations and 6 variables. For outlier contamination data to be included in the original data taken as much as 5% (0.05 x 31 = 2 outliers data) and 10% (0.1 x 31 = 3 outliers data), so that each data is 33 and 34. The contamination data comes from $X \sim N_p (3\mu, \Sigma)$, causes the range of observations of the data
center to be farther away and does not follow the distribution of data patterns so that observations on this data can be called outliers. Contamination data aims to see the effectiveness of methods in detecting outliers.

2.2. Description of Variable
The variables used in this study are Number of murder cases (X1), Number of rape cases (X2), Number of Fraud cases (X3), Number of Persecution cases (X4), Number of cases of damage (X5), and Number of theft cases (X6).

2.3. Mahalanobis Distance
Mahalanobis distance is obtained by calculating the distance of each observation of the data center. The mahalanobis distance square is calculated by the formula (Mahalanobis, 1936) which is defined by the following equation:

$$d_i^2 = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu), i = 1,2, ..., p$$  \hspace{1cm} (1)

Steps to detect outliers with Mahalanobis distance (Johnson, and Wichern, 2007):
1. Determine the average vector value ($\mu$)
2. Determining the value of variance covariance matrix ($\Sigma$)
3. Determine the value of the Mahalanobis distance for each observation with the average vector:

$$d_i^2 = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu), i = 1,2, ..., p$$

4. Sort the value of $d_i^2$ from small to large $d_1^2 \leq d_2^2 \leq \ldots \leq d_n^2$
5. The distance of Mahalanobis is evaluated by using $\chi^2$ on the degree of freedom (df) a number of variables used in the study. Identification of outlier data on the i-observation is defined as outliers if $d_i^2 \geq \chi_{p,1-\alpha}^2$

2.4. Minimum Vector Variance
The Minimum Vector Variance Criteria (MVV) for estimating location and dispersion, first introduced by Herwindiati (2009) by considering data set $X = \{X_1, X_2, \ldots, X_n\}$ from one observation with variables p and $H \subseteq X$. The Vector Variance in the MVV method is the number of squares of the main diagonals found in the sample variance-covariance matrix. The parameters of the MVV estimator are pairs ($x_{mvv}, s_{mvv}$) which are obtained from the smallest Vector Variance. These parameters can be written as follows (Herwindiati, 2009):

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$  \hspace{1cm} (2)
$$s = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^T$$  \hspace{1cm} (3)

The process of calculating the estimated value of the MVV estimator is used the MVV algorithm approach. This algorithm basically calculates objective values from all possible subset of data obtained based on $h = \frac{n(n+1)}{2}$ data with the $s_{mvv}$ variance-covariance matrix that has the minimum $Tr\left(s_{mvv}^2\right)$ value between the set $H$ as much as possible $h$ data. The MVV algorithm for determining parameter estimators is given as follows:
1. Form a $H_{\text{old}}$ subset by taking any observation vector as much as $h = \frac{n + \alpha + 1}{2}$ data from the initial data.
2. Calculate the average vector $\overline{X}(H_{\text{old}})$ and variance covariance matrix $S(H_{\text{old}})$
3. Calculate the value of the mahalanobis distance for all data based on the results in step 2 as follows:
   \[
   MD_{S_{H_{\text{new}}}} = \sqrt{\left( x_{i} - \overline{X}(H_{\text{old}}) \right)^{T} S_{H_{\text{old}}}^{-1} \left( x_{i} - \overline{X}(H_{\text{old}}) \right)}
   \]
4. Sort the value of the mahalanobis distance from the smallest to the largest value such that:
   \[
   MD_{S_{H_{\text{old}}}}(\tau_{1}) \leq MD_{S_{H_{\text{old}}}}(\tau_{2}) \leq \ldots \leq MD_{S_{H_{\text{old}}}}(\tau_{n})
   \]
5. Determine the set $H_{\text{new}}$ as many observations obtained according to step 4 and are defined as:
   \[
   H_{\text{new}} = \{ x_{(\tau_{1})}, x_{(\tau_{2})}, x_{(\tau_{3})}, \ldots, x_{(\tau_{q})} \}
   \]
6. Repeat steps 2 and 3 to get $\overline{X}(H_{\text{new}}), S(H_{\text{new}})$, then $MD_{S_{H_{\text{new}}}}$. If $Tr(S_{H_{\text{new}}}) = Tr(S_{H_{\text{old}}})$ the process is stopped. Conversely, if $Tr(S_{H_{\text{new}}}) < Tr(S_{H_{\text{old}}})$ the process is continued until iterates by repeating steps 2-6. If found $Tr(S_{k}) = Tr(S_{k-1})$, then the iteration process is stopped. Therefore, $\text{Tr}(S_{k}) \geq \text{Tr}(S_{k-1}) \geq \text{Tr}(S_{k-2}) \geq \ldots \geq \text{Tr}(S_{1}) = \text{Tr}(S_{0})$.
7. Robust distance calculation with MVV estimator namely $T_{MVV} = \overline{X}(H_{\text{old}})$ and $S_{MVV} = S(H_{\text{old}})$ which is the estimator pair in calculating robust distance. Robust distance (RdMVV) for each observation vector based on $T_{MVV}$ dan $S_{MVV}$ denoted $\text{RdMVV}(T_{MVV}, S_{MVV})$. So that for robust distance is defined as follows:
   \[
   \text{RdMVV}(T_{MVV}, S_{MVV}) = \sqrt{\left( x_{i} - T_{MVV} \right)^{T} S_{MVV}^{-1} \left( x_{i} - T_{MVV} \right)}
   \]
8. Data outflow is determined if $\text{RdMVV} \geq \chi^{2}_{(\alpha, \nu)}$.

3. Result
3.1 Outcome Detection Result Using Mahalanobis Distance
The results of data processing are in accordance with the Mahalanobis distance analysis procedure using the R program. In calculating the mahalanobis distance value, the average vector and the covariant variant matrix are used. The following is a table of outlier detection results in the original data and data with outlier contamination of 5% and 10% in criminality data that has 6 variables using the Mahalanobis distance:

| Outlier Contamination | Number of Observations(n) | Number of Variables (p) | Number of Outliers Detected | Cut-Off Value $\chi^{2}_{(0.05,6)}$ |
|-----------------------|---------------------------|-------------------------|-----------------------------|-----------------------------------|
| 0%                    | 31                        | 6                       | 4                           | 12.59                             |
| 5%                    | 33                        | 6                       | 5                           | 12.59                             |

Table 1 Results of Outlier Detection of Mahalanobis Distance
Table 1 shows that there are 4 observations classified as outliers in the original data, namely, South Sumatra, Metro Jaya, West Java, and Papua. Meanwhile, in outlier contamination of 5% there are 5 classified outliers, namely, North Sumatra, Metro Jaya, Papua, and 2 outliers of contamination. The 10% contamination data produced 6 outliers, namely, North Sumatra, Metro Jaya, Papua, and 3 outliers of contamination. That is, there is one observation that is detected as outlier in the original data, namely West Java, but not detected when contaminated with outliers.

3.2. Outcome Detection Result Using Minimum Vector Variance

Based on the MVV analysis procedure on multivariate data if the results of $\text{Tr}(\mathbf{S}^2_{new}) = \text{Tr}(\mathbf{S}^2_{old})$ are obtained, the process is stopped. Then, the minimum data with $\text{Tr}(\mathbf{S}^2)$ that will be used to calculate the MVV Distance. The following are the results of the iterations in the original data and data with 5% and 10% outliers contamination:

| Iteration | $\text{Tr}(\mathbf{S}^2)$ Outliers 0% | $\text{Tr}(\mathbf{S}^2)$ Outliers 5% | $\text{Tr}(\mathbf{S}^2)$ Outliers 10% |
|-----------|--------------------------------------|--------------------------------------|--------------------------------------|
| 1         | 4.930243e+12                         | 3.8915e+12                          | 3.846662e+12                        |
| 2         | 41713372554                           | 32465289957                         | 40550054662                         |
| 3         | 41134754773                           | 32465289957                         | 32465289957                         |
| 4         | 45930571624                           | -                                   | 32465289957                         |
| 5         | 45930571624                           | -                                   | -                                   |

Source: Data processed, 2019

Based on Table 1, the minimum $\text{Tr}(\mathbf{S}^2)$ obtained in the original data is in the 3rd iteration, while the lowest $\text{Tr}(\mathbf{S}^2)$ in the 5% and 10% outlier contamination data is in the 2nd and 3rd iterations. Therefore, the data in the most minimum iteration is used to determine the distance of the MVV. The processing results are in accordance with the MVV analysis procedure using the R program. The following is a table of outlier detection results using the MVV method:

| Outlier Contamination | Number of Observations (n) | Number of Variables (p) | Number of Outliers Detected | Cut-Off Value |
|-----------------------|----------------------------|-------------------------|-----------------------------|---------------|
| 0%                    | 31                         | 6                       | 11                          | 12.59          |
| 5%                    | 33                         | 6                       | 13                          | 12.59          |
| 10%                   | 34                         | 6                       | 14                          | 12.59          |
Table 3 shows that there were 11 observations classified as outliers in the original data, namely, North Sumatra, West Sumatra, South Sumatra, Lampung, Metro Jaya, West Java, NTT, West Kalimantan, North Sulawesi, South Sulawesi, and Papua. Whereas, on outbreak contamination of 5% there were 13 observations classified as outliers, namely all observations including outliers in the original data and 2 outliers of contamination. In the 10% contamination data there are 14 observations classified as outliers, namely all observations including outliers in the original data and 3 outliers of contamination. That is, all observations detected as outliers in the original data are also detected when contaminated with outliers.

4. Conclusion

Based on the research that has been done and described in the discussion, it can be concluded that MVV is more effective than the distance of Mahalanobis when detecting outliers with contamination data. Because at the Mahalanobis distance there is 1 observation that is detected as an outlier in the original data but is not detected when contaminated with outliers. Meanwhile, on the MVV method all observations detected as outliers in the original data were also detected in outlier contamination data.

5. Acknowledgement

The author would like to thank the final assignment supervisor with the support of knowledge provided to the author so that this research can be completed.

6. Reference

[1] Rousseeuw P J & Van Zomeren B C 1990 Unmasking Multivariate Outlier and Leverage Points Journal of the American Statistical Association pp85: 633-639.
[2] Noeryanti 2001 Detecting Multiple Outliers in Multivariate Samples With S-Estimation Method Jurnal Teknologi Industri pp 5(4):217 – 226.
[3] Boni Melda Putri 2018 Mengelompokkan Subjek Menggunakan Mahalanobis Distance dan PCA (Sumatra Utara: Tesis Magister Jurusan Matematika Universitas Sumatra Utara)
[4] Herwindiati DE and Sani M 2009 The Robust Principal Component Using Minimum Vector Variance Proceedings of the World Congress on Engineering 2009 Vol I, 1 – 3 July 2009 (London, U.K)
[5] Ali Hazlina 2013 On Robust Mahalanobis Distance issued from minimum vector variance Far East Journal of Mathematical Sciences (Malaysia: Universiti Utara Malaysia)