BAREKENG: Journal of Mathematics and Its Application
September 2022 Volume 16 Issue 3 Page 797–804
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss3pp797-804

OUTLIER DETECTION ON HIGH DIMENSIONAL DATA USING MINIMUM VECTOR VARIANCE (MVV)

Andi Harismahyanti [1] *, Indahwati [2], Anwar Fitrianto [3], Erfiani [4]

[1,2,3,4] *Department of Statistics and Data Science, Faculty of Mathematics and Natural Sciences, IPB University Dramaga St., Campus IPB, Bogor, 16680, West Java, Indonesia*

Corresponding author's e-mail: [1]* *andiharismahyanti@gmail.com*

Abstract. *High-dimensional data can occur in actual cases where the variable p is larger than the number of observations n. The problem that often occurs when adding data dimensions indicates that the data points will approach an outlier. Outliers are parts of observations that do not follow the data distribution pattern and are located far from the data center. The existence of outliers needs to be detected because it can lead to deviations from the analysis results. One of the methods used to detect outliers is the Mahalanobis distance. To obtain a robust Mahalanobis distance, the Minimum Vector Variance (MVV) method is used. This study will compare the MVV method with the classical Mahalanobis distance method in detecting outliers in non-invasive blood glucose level data, both at p>n and n>p. The test results show that the MVV method is better for n>p. MVV shows more effective results in identifying the minimum data group and outlier data points than the classical method.*

Keywords: *outlier detection, high-dimension, Mahalanobis distance, Minimum Vector Variance*

Article information:

Submitted : 29th March 2022 *Accepted : 7th July 2022*

How to cite this article:
A. Harismahyanti , Indahwati, A. Fitrianto and Erfiani , "OUTLIER DETECTION ON HIGH DIMENSIONAL DATA USING MINIMUM VECTOR VARIANCE (MVV)", *BAREKENG: J. Math. & App*., vol. 16, iss. 3, pp. 797-804, September, 2022.

1. INTRODUCTION

Along with the rapid development of technology, the role of data is now crucial in various fields of knowledge and its use. The development increased the number of databases both in the number of observations and in the number of dimensions. In real cases, high-dimensional data can occur where p is greater than n, p is a variable or variable, and n is the number of observations [1]. In theory, increasing the number of variables gives an accurate classification result. However, in practice, with a limited number of observations and a large number of variables, data handling experiences a high analytical complexity in addressing the problems contained in the data [2].

The problem that often occurs in adding data dimensions indicates that the data points will approach an outlier. Outliers are part of the data from a data set that does not follow the data distribution pattern and is located far from the data center. Outliers in the data can result in inaccurate data analysis results, such as deviations from statistical test results based on the mean and covariance parameters [3]. Therefore, detection of outlier indications is needed, especially in extensive data.

Outlier detection is beneficial in various applications such as network intrusion detection, indications of credit card fraud, monitoring activities, financial applications, analysis of election irregularities, bad weather prediction indications, geographic information systems, and other data fields [4]. Detection of outliers is usually with the concept of proximity-based on its relationship to the rest of the existing data. In high-dimensional data, the data density will decrease, resulting in the estimation of the proximity between the data becoming less accurate [5].

One of the methods used to identify outliers in multivariate data is the Mahalanobis distance, by calculating the distance of each observation to the center of the data set [6] . However, the Mahalanobis distance is still included in the classical estimator, which relies on basic assumptions such as normality, linearity, etc. Therefore, a robust method is needed against outliers [3] . One alternative method has been discussed to obtain a robust Mahalanobis distance in detecting outliers in multivariate data, including [3] , [7] , [8], and [9], using the Minimum Vector Variance (MVV) method. However, in this discussion, the multivariate data used is not classified as high-dimensional data with p greater than n. The MVV method utilizes the total variance in finding the minimum covariance matrix so that a robust Mahalanobis distance can be obtained against the outliers. In this study, the outlier detection method will be carried out on multivariate, high-dimensional data.

2. RESEARCH METHODS

The data used in this study are primary data as part of a study by a non-invasive biomarking team at the Bogor Agricultural University regarding the development and clinical trial of a prototype of a non-invasive blood glucose monitoring device. Data collection starts on July 13–20, 2019 using an invasive non-blood glucose measurement tool design. The design captures the intensity of light passed from the finger and was implemented with a total of 74 respondents who came from Kebon Pedes Village, Tanah Sereal District, and Bogor City. This blood glucose level data is generated from intensity residual points from each modulation and time domain. So, before detecting outliers in this data, a summary process is first carried out to obtain the variables that define this blood glucose level.

2.1 Data Summary Process

Statistical analysis of the data is done by calculating the area of the trapezoid to obtain comprehensive information by utilizing the time domain interval set to adjust the point of observation. The following illustrates the data summary process shown in Figure 1.

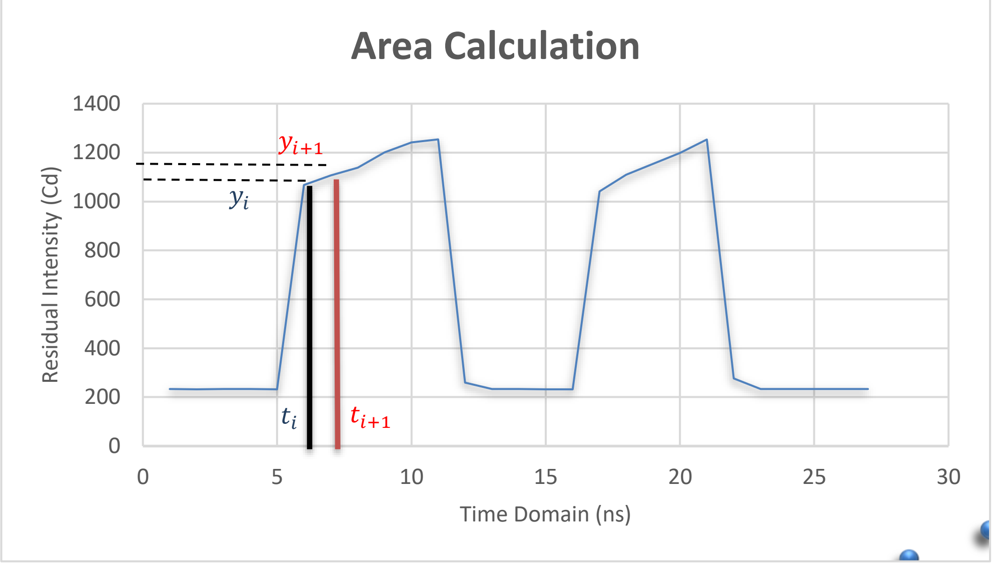

Figure 1. Illustration of Summarizing Data with Peak Area Calculation

The non-invasive measurement of blood glucose levels produces initial data in the form of intensity residues for each time domain which is designed with a lighting level (modulation) of 0-90. Aurelia's research [10] states that broad summarization can estimate blood glucose levels with better performance than standard deviation summaries. This is because area summarization can utilize all the information from the data well and uses broad limits based on time-domain intervals that have been set to suit all observations. Aurelia then discusses using the 50-90 modulation in blood glucose level data in 2017 and 2019, showing significant residual values. Other modulations tend to be constant, so the modulation used in this study is the 50-90 modulation, which is the 26-30 period.

Each residual intensity value in one modulation, as shown in Figure 1, will be drawn in a straight line in the direction of its time domain. Then calculate each area on each peak formed using the trapezoid area formula. The length of the time-domain interval is defined as the height(t_i) and the residual value of the intensity is defined as a parallel side (y_i). After obtaining the area in one modulation, the value of the area is added up to form the peak area of each modulation. One independent variable is the sum of the area values of one modulation.

There were five modulations and five independent variables in one replicate. One independent variable is the sum of the area values of one peak in one modulation. There are two peaks in one modulation, so there are 50 independent variables in five replications. To obtain high-dimensional data with p greater than n, the data will be taken in as many as 30 or 40 observations. The data will be standardized first before conducting further analysis.

Mahalanobis distance calculation is defined by calculating the distance of each observation to the center of all the data. Mahalanobis distance is more practical than Euclidean distance, where the calculation considers the correlation between variables [11] . Multivariate high-dimensional data is very susceptible to the correlation between variables [12] . In this study, the identification of outliers in blood glucose level data in 2019 will be carried out using classical and robust methods.

2.2 Classic Mahalanobis Distance Detection

The steps for detecting outliers using the classical Mahalanobis distance method are as follows [13] :

1. Calculate the average value of each variable vector($\boldsymbol{\mu}$)
2. Calculating the value of the covariance variance matrix from the data set($\boldsymbol{\Sigma}$)
3. Calculate the value of the Mahalanobis distance for each observation point with the average vector with the formula

$$\mathrm{d}_\mathrm{i}^2 = (\mathbf{x}_\mathrm{i} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\mathrm{i} - \bar{\mathbf{x}}), \mathrm{i} = 1,2, \dots \mathrm{p} \quad (1)$$

4. Sort values d_i^2 from the minimum to the largest value $\mathrm{d}_1^2 \leq \mathrm{d}_2^2 \leq \cdots \leq \mathrm{d}_\mathrm{n}^2$

5. Evaluating the Mahalanobis distance using the chi-square cut-off value (χ^2), i.e., if $d_i^2 > \chi^2_{p,(1-\alpha)}$, then the i-th observation point is identified as an outlier.

2.2 Outlier Detection with Mahalanobis Minimum Vector Variance

Generally, there are two methods for dealing with multivariate problems related to covariance: total variance (TV) and general variance (GV). The general variance (GV) is usually called the covariance determinant (CD) or is defined as $|\Sigma|$. The TV role is defined as $\mathrm{Tr}(\Sigma)$, usually used in dimensional reduction problems such as principal component analysis, etc. The role of CD can be used in almost all multivariate problems. However, in its application to the principal component analysis, the CD has limitations if its value is close to zero or equal to zero. Therefore, a new concept for solving this problem was launched, known as vector variance (VV) or defined as $\mathrm{Tr}(\Sigma^2)$ [3].

The effectiveness of VV computation led to the development of VV as a robust estimation by minimizing the vector variance of the data. MVV criteria in labeling outliers and the application of principal components were first introduced by Herwindiati (in [3]) by considering the data set $\mathbf{x}$ = { $x_1, x_2, \ldots, x_n$} from one observation with variables p and H $\subseteq\mathbf{x}$ **.** Let T_{MVV} and S_{MVV} be the MVV estimates for the location parameter and the covariance variance matrix, respectively. Estimates are obtained based on the set H. The number of element locations of H is h = $\frac{(n+p+1)}{2}$ data which will give a covariance variance matrix S_{MVV} with a minimum Tr (S^2_{MVV}) value for all possible sets containing h data. Therefore, the MVV estimates for the location parameters of the matrix are written in the following formula [14];

$$\mathbf{T}_{MVV}=\frac{1}{h}\sum_{i\in H} x_i \tag{2}$$

$$\mathbf{S}_{MVV}=\frac{1}{h-1}\sum_{i\in H}(\mathbf{x}_i - T_{MVV})(\,\mathbf{x}_i - T_{MVV})' \tag{3}$$

The algorithm for detecting outliers with the Minimum Vector Variance (MVV) method is as follows [15]:

1. Takes a data set defined as $\mathbf{H}_{old}$ consisting of h= $\frac{(n+p+1)}{2}$ data.
2. Calculating mean vector $\bar{\mathbf{x}}_{H_{old}}$and covariance matrix $\mathbf{S}_{H_{old}}$for all data $\mathbf{H}_{old}$. Next, for $i = 1,2,\ldots,n$, calculate the Mahalanobis distance with the MVV estimator using the formula.
$$d^2_{H_{old}}= d^2_{H_{old}}(\,\mathbf{x}_i,\bar{\mathbf{x}}_{H_{old}}) = (\mathbf{x}_i - \bar{\mathbf{x}}_{H_{old}})'\,\mathbf{S}^{-1}_{H_{old}}(\mathbf{x}_i - \bar{\mathbf{x}}_{H_{old}}) \tag{4}$$
3. Sort the calculation results $d^2_{H_{old}}$from the smallest to the largest $d^2_{H_{old}}(\,d^2_{H_{old}}(\,\pi_1)\ldots\pi_2)(\,d^2_{H_{old}}.\pi_n)$. This order will give a permutation of the observation index**.**
4. Form a new set $\mathbf{H}_{new}$, It consists of h observations with index $\pi(1), \pi(2),\ldots, \pi(h)$.
5. Counting $\bar{\mathbf{x}}_{H_{new}}$, $\mathbf{S}_{H_{new}}$ and $d^2_{H_{new}}(\,\mathbf{x}_i,\bar{\mathbf{x}}_{H_{new}})$ as in step 2.
6. If Tr($\mathbf{S}^2_{H_{new}}$) < Tr($\mathbf{S}^2_{H_{old}}$) then the process is continued until the kth iteration reaches Tr($\mathbf{S}^2_{H_{new}}$) = Tr($\mathbf{S}^2_{H_{old}}$). At the end of the kth iteration will have Tr($\mathbf{S}^2_{H_1}$) Tr($\mathbf{S}^2_{H_2}$)….≥ Tr($\mathbf{S}^2_{H_{k-1}}$)= Tr($\mathbf{S}^2_{H_k}$).
7. $d^2_{H_k}$ defined as the distance of the Mahalanobis rigid to the outlier. $d^2_{H_k}$ calculated at the end of the selected k-th iteration using the formula
$$d^2_{H_k} = d^2_{MVV}(\,\mathbf{x}_i,\bar{\mathbf{x}}_{MVV}) = (\mathbf{x}_i - \bar{\mathbf{x}}_{MVV})'\,\mathbf{S}^{-1}_{MVV}(\mathbf{x}_i - \bar{\mathbf{x}}_{MVV}) \tag{5}$$
8. Evaluating the Mahalanobis robust distance using a chi-square cut-off value (χ^2), i.e., if $d^2_{MVV} > \chi^2_{p,(1-\alpha)}$, then the observation point is identified as an outlier.

3. RESULTS AND DISCUSSION

3.1. Outline Detection Classical Mahalanobis Distance Method with p>n

The results of outlier detection using the classical Mahalanobis distance method on blood glucose levels are shown in the following Table 1:

Table 1. Outlier Detection with Classic Mahalanobis Distance

Variable (p)	Number of observations (n)	Amount of Withdrawal	Cut-off value$X^2_{(50;0,95)}$
50	30	5	67.50481
50	40	17	67.50481

Table 1 shows the data for non-invasive blood glucose levels in 2019 with some outliers. The distance between Mahalanobis data in 2019 and the number of observations $n = 30$ and $n = 40$ results is between $(-95,22) - 385,67$. Cut - off value$(X^2_{(50;0,95)})$ on the data is $67,50481$. Observations are said to be outliers if they exceed the cut-off limit of the data. Based on the results of the calculation of the Mahalanobis distance, five observations were classified as outliers on $n = 30$, and on $n = 40$ there are 17 outliers. Observations classified as outliers based on the calculation of the Mahalanobis distance can be seen in the scatter plot in Figure 2.

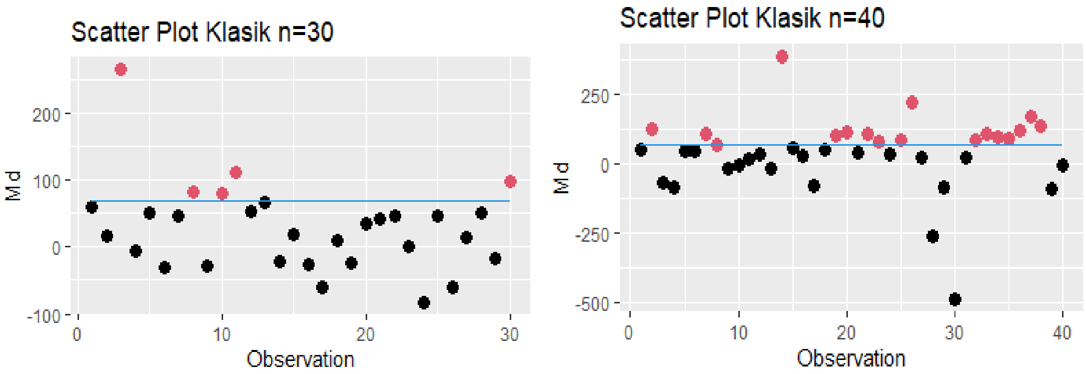

Figure 2. Classic Mahalanobis Distance Plot Scatter

The red dot in Figure 2 indicates outlier observations, while the black dots are normal observations. The horizontal line is the *i-th* observation, and the vertical line Md is the Mahalanobis distance. The blue horizontal line is the cut-off value $(X^2_{(50;0,95)})$, which is 67.50481. The outliers detected tend to be in the area of the blue horizontal line, with the position of the scattered dots making it difficult to identify the group of outliers.

3.2. Outline Detection of Mahalanobis MVV Distance Method with p>n

The results of outlier detection using the Mahalanobis MVV distance method on blood glucose level data are shown in the following Table 2:

Table 2. Detection of Outliers with Mahalanobis Distance MVV

Variable (p)	Number of observations (n)	Amount of Withdrawal	Cut-off value$X^2_{(50;0,95)}$
50	30	4	67.50481
50	40	6	67.50481

The Mahalanobis distance data for 2019 uses the robust MVV method with a large number of observations, $n = 30$and $n = 40$the resulting data ranges from $41,52 - 1896,15$. The cut-off value$(X^2_{(50;0,95)})$ on the data is $67,50481$. Observations are said to be outliers if they exceed the cut-off limit of chi-square, so based on the results of the calculation of the distance of the Mahalanobis Robust MVV, there are four observations classified as outliers in $n = 30$ and $n = 40$, there are six outliers. The number of outliers detected in the Mahalanobis MVV distance method is smaller than in the classical Mahalanobis distance method. Observations classified as outliers based on calculating the distance Mahalanobis MVV can be seen in the scatter plot in Figure 3.

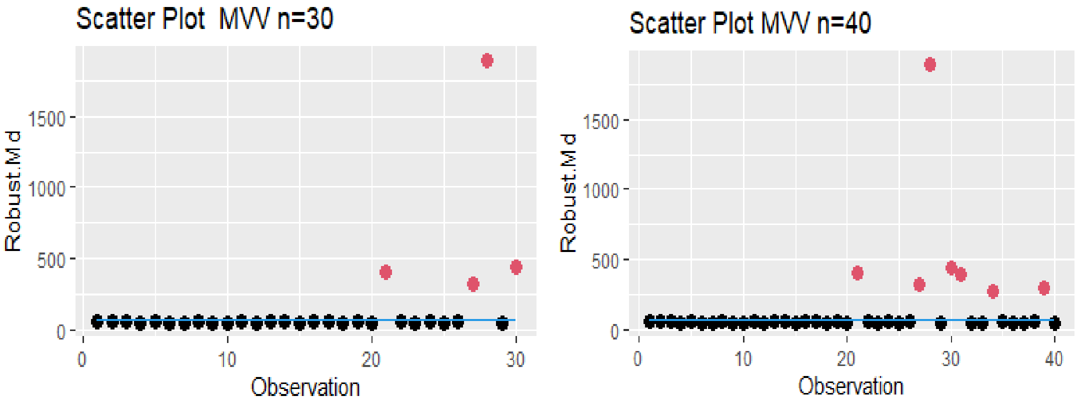

Figure 3. Scatter Plot Mahalanobis Minimum Vector Variance (MVV)

Observation of outliers detected in Figure 3 shows outlier points more clearly grouped in position than the classical method. The robust Mahalanobis distance obtained can identify the minimum data group and indicate data points with extreme outliers.

3.3. Mahalanobis Distance Outage Detection with n>p

Detection of outliers in multivariate data with the number of observations greater than the variable. The data used is data on blood glucose levels with the use of all observations as many as 74 observations. The detection results are summarized in the following table:

Table 3. Detection of outliers with Classical Mahalanobis Distance vs. MVV on n>p. data

Variable (p)	Method	Amount of Withdrawal	Cut-off value$X^2_{(50;0,95)}$
50	Classic	2	67.50481
50	MVV	12	67.50481

The Mahalanobis distance of this data uses the robust MVV method in all observations of the 2019 blood glucose level data produced, ranging from $27,35 - 1896,15$. N value cut-off value ($X^2_{(50;0,95)}$) on the data is $67,50481$. Observations are said to be outliers if they exceed the cut-off limit of the data. Based on the calculation of the Mahalanobis distance results, two observations were classified as outliers, and the robust MVV method identified 12 outliers. The number of outliers detected in the Mahalanobis MVV distance method is greater than in the classical Mahalanobis distance method. This is inversely proportional to the p>n dimension data, where the number of observations identified as outliers by the MVV method is smaller than the classical method. Observations classified as outliers based on the calculation of the Mahalanobis MVV distance can be seen in the scatter plot in Figure 4.

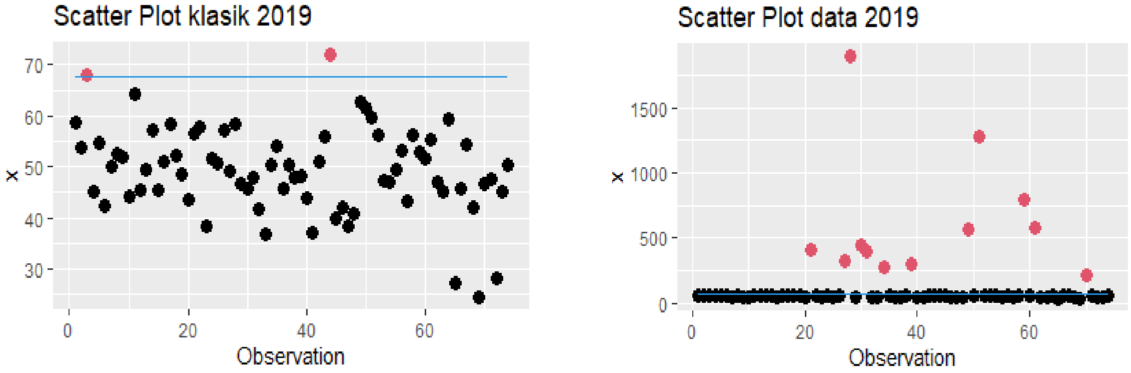

Figure 4. Scatter Plot of Classical Mahalanobis vs. Minimum Vector Variance (MVV) on n>p.

The vertical line labeled x in Figure 4 is defined as the Mahalanobis distance. Some outlier observations are detected by the MVV method but not by the classical method. It is necessary to find points that indicate outliers to facilitate further analysis of blood glucose level data. So, if these points can be detected, it will help in further statistical analysis. Observations of outliers detected in the MVV method show outlier points that are clearer in grouping positions than the classical method on n>p data. The number of outliers obtained from Mahalanobis distance can identify the minimum data group and data points with extreme outliers.

4. CONCLUSION

Based on the study's results, it can be concluded that the Mahalanobis MVV distance is better used for detecting outliers of non-invasive blood glucose level data in 2019 with n>p. The use of the classical Mahalanobis distance is limited in identifying extreme outliers in this data compared to the MVV method, which is more robust against outliers. For high-dimensional data where p>n, the results obtained by the MVV method are more effective in identifying the minimum data group and data points with extreme outliers than the classical Mahalanobis distance method. However, some outliers in the high-dimensional data can be identified at the classical Mahalanobis distance but not identified at the Mahalanobis MVV distance.

REFERENCES

[1] M. Rochayani, “Hybrid Undersampling, Regularization, and Decision Tree Methods for Classification of High Dimensional Data with Unbalanced Classes,” 2020, Accessed: Mar. 27, 2022. [Online]. Available: http://repository.ub.ac.id/183689/.

[2] T. RAHMATIKA, “Support Vector Machine for Multiclass Imbalanced on High Dimensional Data,” 2020, Accessed: Mar. 27, 2022. [Online]. Available: http://etd.repository.ug.ac.id/penelitian/detail/183304.

[3] E. Herdiani, P. Sari, NS-J. of P. Conference, and undefined 2019, “Detection of Outliers in Multivariate Data using Minimum Vector Variance Method,” iopscience.iop.org , doi: 10.1088/1742-6596/1341/9/092004.

[4] E. Wahyuni, SS- Science, undefined technology, undefined Engineering, and undefined 2020, “A Comparison of Outlier Detection Techniques in Data Mining,” seminar.uad.ac.id , Accessed: Mar. 26, 2022. [Online]. Available: http://seminar.uad.ac.id/index.php/STEEEM/article/download/2878/805.

[5] GN-J. of AI System and undefined 2016, “Detection of Transaction Outliers Using Visualization-Olap in Private Higher Education Data Warehouses,” publications.dinus.ac.id , Accessed: Mar. 26, 2022. [Online]. Available: http://publikasi.dinus.ac.id/index.php/jais/article/view/1184.

[6] J. Mei, M. Liu, Y. Wang, HG-I. transactions on, and undefined 2015, “Learning a Mahalanobis distance-based dynamic time warping measure for multivariate time series classification,” ieeexplore.ieee.org , Accessed: Mar. 28, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7104107/.

[7] M. FARUK, "Comparison of MVV and FMCD methods in detecting outliers in a normal multivariate data observation," 2008, Accessed: Mar. 27, 2022. [Online]. Available: http://etd.repository.ugm.ac.id/home/detail_pencarian/39086.

[8] D. Juniardi, MM-BBI Mathematics, and undefined Statistika, “USE OF MINIMUM VECTOR VARIANCE (MVV) METHOD AND CONFIRMATION ANALYSIS IN DETECTING OUTLIER,” journal.untan.ac.id , vol. 01, no. 1, pp. 31–40, 2012, Accessed: Mar. 25, 2022. [Online]. Available: https://jurnal.untan.ac.id/index.php/jbmstr/article/view/5187.

[9] Juniardi DKMNM, “USE OF MINIMUM VECTOR VARIANCE (MVV) METHOD AND CONFIRMATION ANALYSIS IN DETECTING OUTLIER,” Bimaster Bul. science. Matt. stats. and Ter. , vol. 3, no. 01, March. 2014, doi:10.26418/BBIMST.V3I01.5187.

[10] K. Aurelia, “Non-invasive Estimation of Blood Glucose Levels Using Partial Least Square Regression with Multiple Summary Approaches,” 2020, [Online]. Available: https://repository.ipb.ac.id/handle/123456789/104399.

[11] MP Boni et al. , “Mahalanobis Distance And Pca,” 2018.

[12] C. Leys, O. Klein, Y. Dominicy, and C. Ley, “Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance,” J. Exp. soc. Psychol. , vol. 74, pp. 150–156, Jan. 2018, doi:10.1016/J.JESP.2017.09.011.

[13] R. Johnson, DW- Statistics, and undefined 2015, “Applied multivariate statistical analysis,” statistics.columbian.gwu.edu, Accessed: Mar. 28, 2022. [Online]. Available: https://statistics.columbian.gwu.edu/sites/g/files/zaxdzs1911/f/downloads/Syllabus Stat 6215.G Wang Fall 2015.pdf.

[14] DE Herwindiati and SM Isa, “The Robust Principal Component Using Minimum Vector Variance,” Proc. World Congr. eng. , vol. 1, pp. 325–329, 2009.

[15] N. Mukhtar, “ANALYSIS OF MAIN COMPONENTS OF ROBUST USING MINIMUM VECTOR VARIANCE METHOD NURHARDIANTI MUKHTAR'S thesis,” no. April, 2019.