Minimum Vector Variance Estimator in Outlier labeling of Multivariate Data: Application to HIV patient in Indonesia

Erna Tri Herdiani[1], Nurtiti Sunusi[1*], and Puji Puspa Sari[1]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, South Sulawesi, Indonesia, 90245
[*] Corresponding author. E-mail: nurtitisunusi@unhas.ac.id

Received: Oct. 07, 2020; Accepted: May. 17, 2021

An outlier is an observation whose pattern does not follow the majority of the data. Outliers in this study were characterized by extreme distance values, both very small and very large, exceeding the predetermined value. The method used in this research is Minimum Vector Variance (MVV) method because it has good computational efficiency and is robust against outliers. Based on the MVV algorithm applied to data on HIV patients in Indonesia in 2016-2018. The results showed that the MVV method produced more extreme distances than the Mahalanobis distance in labeling outliers. In the research data, it is found that there are 16 regions including outliers of the 34 observation used.

Keywords: HIV, MVV, Outliers, Mahalanobis Distance
© The Author('s). This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202202_25(1).0002

1. Introduction

Detection of outliers on multivariate data is a challenging science for researchers. This is because outlier detection in multivariate data is not easy. The problem becomes complex when there are two or more outliers derived from more than two variables [1]. This difficulty increases when the data is large, which is characterized by a large number of samples and variables. For problems like this, [2] have stated that the efficiency of the calculation is an important thing, as is the effectiveness of the detection process. The existence of outliers in the data can also lead to deviations from the results of data analysis, such as deviations from the results of statistical tests based on average and covariance parameters. Therefore, it is necessary to identify its existence [3].

There are several methods used to detect outliers including Minimum Volume Ellipsoid (MVE) which is based on the smallest ellipsoid volume estimator which includes h from n observations [4], Robust Mahalanobis [1], Minimum Covariance Determinant (MCD) is used to detect outliers based on the determinant value of the variance-covariance matrix [5], Fast Minimum Covariance Determinant (FMCD) is a modification of the MCD method which uses a minimum determinant value of the variance-covariance matrix [6], and the Minimum Vector Variance (MVV) method which is a modification of the Fast MCD (FMCD) algorithm by using a minimum size Vector Variance (VV) [7]. Among these methods, MVV is a more effective method with a lower level of complexity. In addition, according to [8]MVV has excellent properties as an estimator, namely having a high breakdown point value, having an affine equivariance character, and a very high level of computational efficiency. Other research [9] states that the MVV method is robust against data that has been contaminated with outliers. So, this method is good for outliers labeling on multivariate data.

In Indonesia, the number of HIV sufferers has fluctuated in the last three years. Based on data from the Indonesian Ministry of Health [10], the number of HIV sufferers in 2016 was 41,250 cases, then in 2017 it increased to 48,300 cases, and in 2018 the number decreased to 46,659 cases. This puts Indonesia in the top three of the spread of HIV among

Asia Pacific countries. Based on the research conducted [11] it is necessary to increase and expand HIV prevention efforts.

Therefore, support from all sectors is needed. In addition, collective action that is not limited to both the government sector and society. Only in this way can the spread of the HIV epidemic in Indonesia be prevented. This prevention effort is expected to prevent the emergency of new HIV cases. Another study conducted by [12] produced findings that indicate the need for interventions that reduce the impact of HIV stigma on PLHIV. According to [13] people living with HIV also have the potential to contract other diseases that will exacerbate their negative risk. This condition underlies the importance of conducting research on data on HIV sufferers in Indonesia by labeling outliers using the MVV algorithm. Outliers of HIV data in Indonesia are marked with either very small or very large extreme values.

2. Methodology

2.1. Datasets

The multivariate data used in this study comes from the Ministry of Health of the Republic of Indonesia in 2018 [10] regarding data on the number of HIV sufferers in Indonesia for 2016-2018. The research data consisted of 3 variables and 34 observations consisting of all provinces in Indonesia. To show the strength and level of accuracy of the MVV algorithm, contamination data is used. The contamination data comes from $X \sim N_p(3\mu, \sum)$, causing the range of observations to be further from the data center and does not follow the distribution of data patterns so that observations on this data can be categorized as outliers. Contamination data aims to see the effectiveness of the method in detecting outliers. The contamination data consisted of 0%, 5%, and 15% outliers.

2.2. Mahalanobis Distance

Mahalanobis distance is a method for detecting outliers on multivariate data. Mahalanobis distance is obtained by calculating the distance of each observation to the data center. Mahalanobis distance squares are calculated using the formula [14] as follows:

$$d_{MD}^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) > \chi^2_{p,(1-\alpha)} \quad (1)$$

Where $d_i^2 =$ square the distance of observation - i and x_i= observation value -i with,

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix},$$

$$\text{and } \boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix}$$

Steps to detect outliers with Mahalanobis distance [14] :

1. Determine the value of mean vector ($\bar{x}$)
2. Determine the value of the variance covariance matrix ($\boldsymbol{S}$)
3. Determine the value of Mahalanobis distance for every observation with mean vector and variance covariance matrix: $d_i^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}), i = 1, 2, \dots, p$
4. Sort the value d_i^2 from small to large $d_1^2 \leq d_2^2 \leq ... \leq d_n^2$.

Mahalanobis distance is evaluated using χ^2 in the degrees of freedom (df) equal to the number of variables used in the study. This identification of outliers in the observation-i are defined as outliers if the square of Mahalanobis distance is greater than the chi-square value (χ^2) on the p variable [14].

2.3. Algorithm Minimum Vector Variance

Minimum Vector Variance inspired by the C-Steps algorithm in FMCD. Rousseuw and Van Driessen introduced the FMCD method to detect outliers based on the determinant value of the minimum variance-covariance matrix. However, the FMCD method has a weakness when the determinant value of the variance-covariance matrix is zero. [15] modified the FMCD algorithm to be more effective than the previous method and had a lower level of complexity by using a minimum size vector variance (VV). The result of this modification is known as Minimum Vector Variance (MVV).

The MVV criteria for estimates location and dispersion, were first introduced by Herwindiati, 2006 [7] by considering the data set $\boldsymbol{X} = \{X_1, X_2, ...X_n\}$ from an observation with the number of variables is p. Let T_{MVV} and S_{MVV} is the MVV estimate for the location parameter and the variance covariance matrix.

Estimates are obtained based on the set $\boldsymbol{H}$, Where $\boldsymbol{H} \subseteq \boldsymbol{X}$. The number of element locations of $\boldsymbol{H}$ is h = $\frac{(n+p+1)}{2}$ data which will provide a variance covariance matrix S_{MVV} with a minimum $tr(S^2_{MVV})$ value for all possible sets containing h data. Hence, the estimated MVV for the location parameter of the matrix is as follows:

$$T_{MVV} = \frac{1}{h} \sum_{i \in H} X_i \quad (2)$$

Table 1. Result of MVV Data Iteration

Iteration	$Tr(S^2)$ (Outliers 0%)	$Tr(S^2)$ (Outliers 5%)	$Tr(S^2)$ (Outliers 15%)
1	3.281136e+13	9.810788e+13	2.285931e+13
2	40912894560	41988881807	53459115753
3	**39308044790**	40947408334	**38572524955**
4	**39308044790**	**25330500267**	38572524955
5	-	**25330500267**	-

Source: Processed data, 2020

$$S_{MVV} = \frac{1}{h-1} \sum_{i \in H} (X_i - T_{MVV})(X_i - T_{MVV})' \quad (3)$$

The calculation of the estimated value of the MVV estimator is done using the MVV algorithm approach. This algorithm basically calculates the objective value of all possible data subsets obtained based on $h = \frac{n+p+1}{2}$ data, with the variance covariance matrix S_{MVV} which has the minimum value of $tr(S^2_{MVV})$ In determining the estimator of the parameters, the MVV algorithm is used as follows:

1. Taking the data set consisting of $h = \frac{n+p+1}{2}$ data declare this data set with H_{old}.
2. Calculation the mean vector $\overline{X}_{H_{old}}$ and the variance covariance matrix $S_{H_{old}}$ for all data H_{old}. Next for i = 1,2,..., n, determine
   $d^2_{H_{\text{old}}} = d^2_{H_{\text{old}}}(X_i, \bar{X}_{H_{\text{old}}})$
   $= (X_i, -\bar{X}_{H_{\text{old}}})' S^{-1}_{H_{\text{old}}} (X_i, -\bar{X}_{H_{\text{old}}})$
3. Sort the calculation results from smallest to largest. This sequence will give the permutation of the observed index π. For example, the result of sorting the data are
   $d^2_{H_{\text{old}}}(\pi_1) \leq d^2_{H_{\text{old}}}(\pi_2) \leq \cdots \leq d^2_{H_{\text{old}}}(\pi_n)$
4. Form a new set consisting of h observations with index $\pi(1), \pi(2), ..., \pi(n)$, then name it H_{new}.
5. calculating $\bar{X}_{H_{\text{new}}}, S_{H_{\text{new}}}$ dan $d^2_{H_{\text{new}}}(X_i, \bar{X}_{H_{\text{new}}})$ as in the step 2
6. If $Tr(S^2_{H_{new}}) = Tr(S^2_{H_{old}})$ then the process is done. If $Tr(S^2_{H_{new}}) < Tr(S^2_{H_{old}})$ then the process continuous until k-iteration reaches $Tr(S^2_{H_{new}}) = Tr(S^2_{H_{old}})$
7. If S_{H_i} is the variance covariance matrix of the iteration-k. Then at the end of the iteration the k will have $\text{Tr}\left(S^2_{H_1}\right) \geq \text{Tr}\left(S^2_{H_2}\right) \geq \cdots \geq \text{Tr}\left(S^2_{H_{k-1}}\right) = \text{Tr}\left(S^2_{H_k}\right)$
8. If the value of minimum trace has been obtained on the variance covariance matrix, the next step is to sort the multivariate data using the Mahalanobis distance.
9. Mahalanobis distance from the minimum trace is called $Robust\ distance_{MVV}$ which is then evaluated using χ^2 in degrees of freedom (df) as many as the number of variables used in the study. Data identification in the i TH observation is defined as an outlier if $Rd_{MVV} \geq \chi^2_{(p:1-\alpha)}$.

2.4. Breakdown Point

According to [16] the breakdown point is the smallest fraction or percentage of outliers that causes the value of the estimator to be large. The breakdown point is used to explain the breakdown size of the robust technique. The highest possible breakdown point for an estimator is 50%. If the breakdown point is more than 50%, it means that the regression model estimate cannot describe the information from the majority of the data. One of the estimators that has a high breakdown point value is the MVV algorithm [7].

Herwindiati has proven that the breakdown point value of the MVV algorithm is 50% or asymptotically it is 1/2. That is, MVV has a minimal proportion of the number of outliers compared to all observational data. In other words, the MVV algorithm is robust or robust against outliers. Meanwhile, Mahalanobis Distance in this study has a breakdown point value of 0 because outliers affect the mean and standard deviation.

In this study, the results of the Mahalanobis distance are shown not as a comparison of the MVV algorithm, but only shown because in multivariate data, sequencing the MVV algorithm uses the Mahalanobis distance. So, it can be seen the work process when Mahalanobis Distance is used in multivariate data sorting.

3. Result

In this research, the first step is to find the minimum iteration so that the value can be processed in the Minimum Vector Variance algorithm. To find the minimum iteration, the step taken is to find the value of the variance covariance matrix, that squaring the main diagonal value in the variance covariance matrix to obtain the trace value. Then,

Table 2. Outlier Detection Result using the MVV

Outlier Contamination	Number of Observation (n)	Number of variable (p)	Number of Outliers
0%	34	3	16
5%	36	3	18
15%	39	3	21

Source: Processed data, 2020

the trace value that have been obtained are squared. If the result is that $Tr(S^2_{H_{new}}) = Tr(S^2_{H_{old}})$ then the process is terminated, meaning $Tr(S^2)$ the minimum has been found. Then, the data with minimum $Tr(S^2)$ will be used to calculate the Robust MVV distance. The following are the iterations of data on HIV patients in Indonesia with outlier contamination of 0%, 5%, and 15%:

Based on Table 1, it can be seen that the outlier contamination of 0% and 15% obtained $Tr(S^2_{H_{new}}) = Tr(S^2_{H_{old}})$ or the square trace value The minimum is obtained in the 4th iteration because the value in the iteration is the same as the value in the 3rd iteration, while for outlier contamination of 10% the minimum trace value is in the 5th iteration because the value in the iteration is the same as the value in the previous iteration. Since the minimum iteration has been obtained, the next step is to calculate the robust MVV distance using the mean and variance-covariance matrix. The results of data processing are in accordance with the MVV analysis procedure using R. Table 2 shows the results of outlier detection using the MVV method on data with outlier contamination of 0%, 5%, and 15%:

To see the Mahalanobis distance working process in multivariate data sorting, outliers labeling using Mahalanobis Distance and MVV is shown in Table 3. Based on the Table 3, it can be seen that the outlier labeling using the mahalanobis distance resulted in 5 outliers in the data on HIV patients for 2016-2018. The areas that are included in outliers are DKI Jakarta, West Java, Central Java, East Java, and Papua with an average distance value generated of 14.85644. Meanwhile, the outliers labeling using the Minimum Vector Variance algorithm resulted in more outliers because the steps were more complex than the mahalanobis distance. As mentioned by [7, 9] that the MVV method is robust for outliers or robust on data that has been contaminated with outliers.

This means that the results of the labeling used on HIV data are reliable. The areas labeled as outliers using the MVV method are North Sumatra, Riau, Riau Islands, DKI Jakarta, West Java, Central Ja¬va, Yogyakarta Yogyakarta, East Java, Bali, East Nusa Tenggara, South Kalimantan, East Kalimantan, South Sulawesi, Maluku, West Papua and Papua. In Table 3, the regional data shown is data without outlier contamination. Below is a visualization of the area in Indonesia:

Fig. 1, it can be seen that the most dominant areas of HIV spread in 2016-2018 were Java and Papua, while Sumatra, Kalimantan, Bali, Nusa Tenggara and Sulawesi tended to be excluded from outliers.

On Fig. 2 obtained a very complex difference compared to Fig. 1. Fig. 2 is the visualization of outliers labeling using the minimum vector-variance method. Based on this figure, if combined with the results of the calculations in Table 1, it can be seen that the MVV method produces extreme distances so that it is very clear the difference between areas including outliers and non-outliers. Almost every island in Indonesia has an area that is classified as an outcast in the spread of HIV. One of the factors that underlie this is the complexity of calculating the MVV method compared to the Mahalanobis distance.

4. Conclusion

Based on the research that has been done, it can be concluded that the Minimum Vector Variance algorithm produces more extreme when using multivariate data sequencing of Mahalanobis distance in labeling outliers. In the 2016-2018 HIV data in Indonesia, it was found that there were 16 regions including outliers of the 34 observations used. These areas including outliers can become the government's main focus in preventing the spread of HIV in Indonesia.

Acknowledgments

Thank you to the funders through the Higher Education Excellence Basic Research grant, the Ministry of Education and Culture of the Republic of Indonesia.

References

[1] P. J. Rousseeuw and B. C. Van Zomeren, (1990) *"Unmasking multivariate outliers and leverage points"* **Journal of the American Statistical association** *85*(411): 633–639. DOI: 10.1080/01621459.1990.10474920.

[2] F. Angiulli and C. Pizzuti, (2005) *"Outlier mining in large high-dimensional data sets"* **IEEE transactions on Knowledge and Data engineering** *17*(2): 203–215. DOI: 10.1109/TKDE.2005.31.

Table 3. Outliers of Mahalanobis and MVV based on the distribution of HIV in Indonesia

No	Provinsi	Mahalanobis	Outlier Mahalanobis	MVV	Outlier MVV
1	Aceh	0.53873708	No Outlier	1.0018192	No Outlier
2	North Sumatra	3.0308964	No Outlier	**55.2064912**	Outlier
3	West Sumatra	0.47032982	No Outlier	3.3255137	No Outlier
4	Riau	1.2081002	No Outlier	**34.9413544**	Outlier
5	Jambi	0.31502998	No Outlier	0.8453527	No Outlier
6	South Sumatra	0.35581944	No Outlier	1.2345791	No Outlier
7	Bengkulu	0.51159566	No Outlier	1.6806615	No Outlier
8	Lampung	0.57713557	No Outlier	5.9762064	No Outlier
9	Bangka Belitung Islands	0.62793597	No Outlier	4.959546	No Outlier
10	Riau Islands	0.22083396	No Outlier	**11.2631328**	Outlier
11	DKI Jakarta	**16.03054572**	Outlier	**433.6551319**	Outlier
12	West Java	**13.79123729**	Outlier	**477.2490035**	Outlier
13	Central Java	**12.27119696**	Outlier	**340.8644335**	Outlier
14	DI Yogyakarta	1.40524883	No Outlier	**14.1856788**	Outlier
15	East Java	**17.96460748**	Outlier	**687.8867598**	Outlier
16	Banten	0.09978532	No Outlier	7.5102819	No Outlier
17	Bali	2.25911188	No Outlier	**88.1782887**	Outlier
18	West Nusa Tenggara	0.3361526	No Outlier	0.8929734	No Outlier
19	East Nusa Tenggara	1.35816905	No Outlier	**19.4510167**	Outlier
20	West Kalimantan	0.55704081	No Outlier	3.5470756	No Outlier
21	Central Kalimantan	0.4856821	No Outlier	1.8377174	No Outlier
22	South Borneo	1.74180534	No Outlier	**35.4630182**	Outlier
23	East Kalimantan	1.4026929	No Outlier	**23.7734999**	Outlier
24	North Kalimantan	0.39134516	No Outlier	0.8838721	No Outlier
25	North Sulawesi	0.32873171	No Outlier	0.7115748	No Outlier
26	Central Sulawesi	0.73404337	No Outlier	3.185359	No Outlier
27	South Sulawesi	1.8313649	No Outlier	**34.1083183**	Outlier
28	Southeast Sulawesi	0.38661508	No Outlier	1.6459139	No Outlier
29	Gorontalo	0.52024698	No Outlier	0.9923645	No Outlier
30	West Sulawesi	0.45227347	No Outlier	1.1583853	No Outlier
31	Maluku	0.88048809	No Outlier	**24.8359743**	Outlier
32	North Maluku	0.57721752	No Outlier	1.347671	No Outlier
33	West Papua	1.11337752	No Outlier	**16.2766244**	Outlier
34	Papua	**14.22460583**	Outlier	**392.4121644**	Outlier

Source: Processed data, 2020

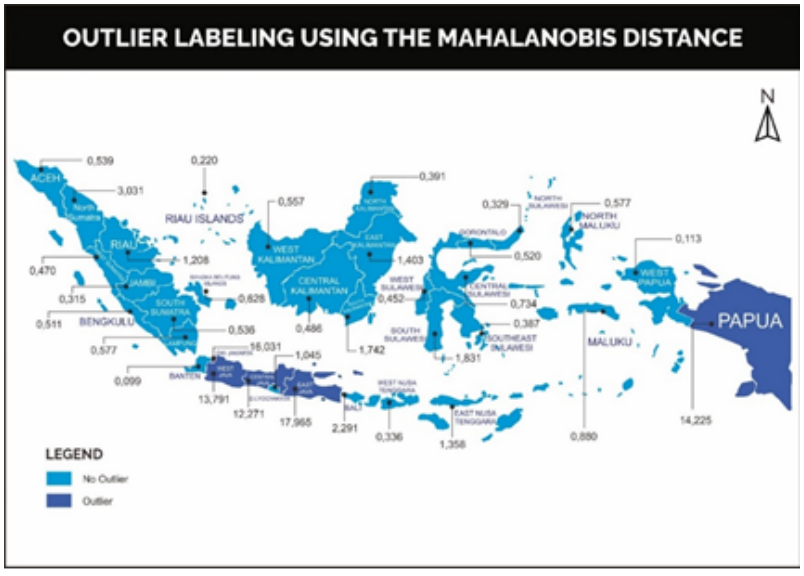

Fig. 1. Outlier Labeling Using the Mahalanobis Distance

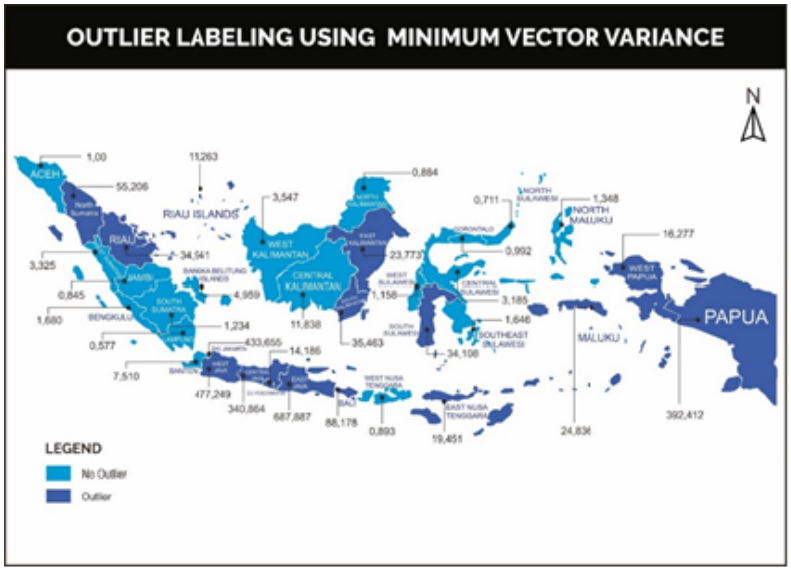

Fig. 2. Outlier Labeling Using Minimum Vector Variance

[3] A. S. Hadi, (1992) *"Identifying multiple outliers in multivariate data"* **Journal of the Royal Statistical Society: Series B (Methodological)** ***54***(3): 761–771. DOI: 10.2307/2345856.

[4] P. J. Rousseeuw and S. Van Aelst, (2009) *"Minimum volume ellipsoid"* **Wiley Interdisciplinary Reviews: Computational Statistics** ***1***: 71–82. DOI: 10.1002/wics.19.

[5] M. Hubert, M. Debruyne, and P. J. Rousseeuw, (2018) *"Minimum covariance determinant and extensions"* **Wiley Interdisciplinary Reviews: Computational Statistics** ***10***(3): e1421. DOI: doi:10.1002/wics.1421.

[6] P. J. Rousseeuw and K. V. Driessen, (1999) *"A fast algorithm for the minimum covariance determinant estimator"* **Technometrics** ***41***(3): 212–223. DOI: 10.1080/00401706.1999.10485670.

[7] D. E. Herwindiati, M. A. Djauhari, and M. Mashuri, (2007) *"Robust multivariate outlier labeling"* **Communications in Statistics—Simulation and Computation®** ***36***(6): 1287–1294. DOI: 10.1080/03610910701569044.

[8] H. Ali, S. S. S. Yahaya, and Z. Omar. "The efficiency of reweighted minimum vector variance". In: *AIP Conference Proceedings*. ***1602***. 1. American Institute of Physics, 2014, 1151–1156. DOI: doi:10.1063/1.4882629.

[9] E. T. Herdiani, P. P. Sari, and N. Sunusi. "Detection of Outliers in Multivariate Data using Minimum Vector Variance Method". In: *Journal of Physics: Conference Series*. ***1341***. 9. IOP Publishing, 2019, 92004. DOI: 10.1088/1742-6596/1341/9/092004.

[10] K. K. R. Indonesia, (2018) *"Kementrian Kesehatan Republik Indonesia"* **Data dan Informasi Profil Kesehatan Indonesia**:

[11] P. Riono and S. Jazant, (2004) *"The current situation of the HIV/AIDS epidemic in Indonesia"* **AIDS education and prevention** ***16***(Supplement A): 78–90. DOI: 10.1521/aeap.16.3.5.78.35531.

[12] G. J. Culbert, V. A. Earnshaw, N. M. S. Wulanyani, M. P. Wegman, A. Waluyo, and F. L. Altice, (2015) *"Correlates and experiences of HIV stigma in prisoners living with HIV in Indonesia: a mixed-method analysis"* **Journal of the Association of Nurses in AIDS Care** ***26***(6): 743–757. DOI: 10.1016/j.jana.2015.07.006.

[13] S. M. Kimani, M. S. Painschab, M.-J. Horner, M. Muchengeti, Y. Fedoriw, M. S. Shiels, and S. Gopal, (2020) *"Epidemiology of haematological malignancies in people living with HIV"* **The Lancet HIV**: DOI: 10.1016/S2352-3018(20)30118-1.

[14] R. Johnson and D. W. Wichern, (2007) *"Applied Multivariate Sttistical Analysis"* **6th edition.New Jersey: Printice Hall**:

[15] D. E. Herwindiati and S. M. Isa. "The robust principal component using minimum vector variance". In: *Proceedings of the World Congress on Engineering*. ***1***. 2009, 325–329.

[16] P. Huber, (1981) *"Robust Statistics"* **Canada: A John Wiley & Sons, Inc.**