Jurnal Varian
https://journal.universitasbumigora.ac.id/index.php/Varian
e-ISSN 2581-2017
Vol. 4, No. 1, Oktober 2020, Hal. 19-24

Robust Singular Value Decomposition Method on Minor Outlier Data

Bernadhita Herindri S. Utami[1], Trisnawati[2], Rani Pratiwi[3], Miswan Gumanti[4]
[1]Information System, STMIK Pringsewu, Indonesia, e-mail : ind.indri1245@gmail.com
[2]Information System, STMIK Pringsewu, Indonesia, e-mail : trisnawatistmikpsw@gmail.com
[3]Information System, STMIK Pringsewu, Indonesia, e-mail : ranipratiwi.sh@gmail.com
[4]Information System, STMIK Pringsewu, Indonesia, e-mail : mgumanti0205@gmail.com

Article history:
Received : 26-08-2020
Revised : 08-09-2020
Accepted : 09-09-2020

Keywords:
Outlier;
Minor outlier;
Robust;
Singular value;
Decomposition

ABSTRACT

In multivariate statistics, Singular Value Decomposition (SVD) for a data matrix containing outliers does not provide data that can be analyzed optimally. This study aims to overcome outlier data using the Robust Singular Value Decomposition (RSVD) method and compare it with the SVD method. The analysis using the RSVD method includes several steps, namely determining the initial predictive value of the vector **u** and regressing it then normalizing the estimator vector **β** and carrying out the iteration process until convergent results are obtained. The results of this study indicate that the RSVD for dealing with minor outliers data is not influenced by initial estimates. The RSVD method is strongly influenced by the large amount of outliers data, the more extreme outliers data, the more iterations are.

DOI: https://doi.org/10.30812/varian.v4i1.857

A. INTRODUCTION

Outline data is a datum that deviates from another set of datum, (Neter, J., Wasserman, W., and Kutner, 1990). To identify outliers data, a scatter plot or boxplot can be used in the statistics software package. In the regression method, the existence of outliers data will interfere with the fulfillment of assumptions so that the resulting model is unreliable. Likewise in the multivariate case, the result is inaccurate interpretations and errors in decision making on the model obtained. That is why outlier datum omitted as much as possible in the data (Liu, Hawkins, Gosh, & Young, 2003). For example, studies accommodating missing data are in mortality data by (Zhang, L., Shen, H., Huang, 2013). The singular value decomposition method was introduced in (Zhang, L., Marron, J.S., Shen, H., 2007)which is used to make sequential estimates of the eigenvalues and left and right eigenvectors and ignore the missing values and are resistant to outliers. In multivariate statistics, outlier data can be overcome by using the Singular Value Decomposition (SVD) method even though it often does not provide the expected results or there are still deviations in the data (Huber & Ronchetti, 2011). For this reason, a better method is needed, in this case, the author proposes the Robust Singular Value Decomposition (RSVD) method as a solution for handling outliers data. The RSVD method is formulated to minimize problems caused by eliminating outliers data through a matrix approach from the data (Liu et al., 2003). The RSVD method utilizes a regression approach by conducting an iteration process for each of the eigenvalues and eigenvectors. This study aims to obtain the estimated results of the RSVD method and compare it with the SVD method.

B. LITERATURE REVIEW

1. Eigenvalue and Eigenvector

If $\boldsymbol{A}$ is an $n \times n$ matrix then the nonzero vector $\boldsymbol{x}$ in $\Re^n$ is called the eigenvector of $\boldsymbol{A}$ is the scalar multiple of $\boldsymbol{x}$, is an equation (1) (Aa, Morsche, & Mattheij, 2007):

$$\boldsymbol{Ax} = \lambda \boldsymbol{x}, \tag{1}$$

for a scalar, λ is called the eigenvalue of $\boldsymbol{A}$ while $\boldsymbol{x}$ is the eigenvector corresponding to λ (Anton, 1987).

2. Singular Value

The singular value of the matrix $\boldsymbol{A}_{n\times n}$ is the root of the eigenvalues of $n \times n$ symmetric matrix $\boldsymbol{A}^T\boldsymbol{A}$ is denoted by σ_1,σ_2,$\cdots$,σ_n and arranged in the order of $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ (Locantore et al., 1999).

3. Singular Value Decomposition

Singular Value Decomposition is a method that can be applied to any matrix of size $m \times n$. This method can also be applied to matrices that have an inverse or not with a matrix with rank = n or rank $< n$. Suppose that $\boldsymbol{X}$ matrix of size $m \times n$ decomposed is an equation (2) (Valverde-albacete & J, 2020)

$$\boldsymbol{X} = \boldsymbol{U} \times \boldsymbol{L} \times \boldsymbol{A}^T, \tag{2}$$

with:

$\boldsymbol{U}$ = matrix of size $m \times k$ and orthogonal (column $\boldsymbol{U}$ is the eigenvector of $\boldsymbol{AA}^T$)

$\boldsymbol{A}$ = matrix of size $k \times n$ and orthogonal (column $\boldsymbol{A}$ is the eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$)

$\boldsymbol{L}$ = diagonal matrix of size $n \times n$ with non-negative diagonal elements which is called a singular value with $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \cdots \geq \sqrt{\lambda_n}$ then the singular value of $\boldsymbol{X}$ is $\sigma_j = \sqrt{\lambda_j}$(Bretscher, 1997).

The following is given the theorem about SVD:

Theorem 2.1 *Let* $\boldsymbol{X} \in \Re^{n\times m}$ *has a rank* r *with* $\boldsymbol{U} \in \Re^{n\times m}$, $\boldsymbol{L} \in \Re^{n\times n}$, $\boldsymbol{A} \in \Re^{m\times m}$ *where* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *is orthogonal*

$$\boldsymbol{L} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & \cdots & 0 \\ 0 & 0 & 0 & \sigma_r & \cdots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$ and $\boldsymbol{X} = \boldsymbol{U} \times \boldsymbol{L} \times \boldsymbol{A}^T$.

Proof:

Suppose $\boldsymbol{a}_1$, $\boldsymbol{a}_2$,$\cdots$, $\boldsymbol{a}_m$ and $\boldsymbol{u}_1$,$\boldsymbol{u}_2$,$\cdots\boldsymbol{u}_n$ are eigenvectors and let σ_1,σ_2,$\cdots$,σ_r be nonzero eigenvalues at $\boldsymbol{X}$ and suppose $\boldsymbol{U}$ and $\boldsymbol{A}$ is orthogonal

$$\boldsymbol{Xa}_i = \begin{cases} \sigma_i \boldsymbol{u}_i\,, \ i = 1,2,\cdots,r \\ 0, \ i = r+1 \end{cases}$$

expressed in matrix form

$$\boldsymbol{Xa}_i = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & \cdots & 0 \\ 0 & 0 & 0 & \sigma_r & \cdots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

then $\boldsymbol{XA} = \boldsymbol{UL}$ where $\boldsymbol{AA}^T = \boldsymbol{I}$ so that

$$\boldsymbol{XA} = \boldsymbol{UL}$$
$$\boldsymbol{XAA^T} = \boldsymbol{ULA^T}$$
$$\boldsymbol{X} = \boldsymbol{ULA^T}$$

4. Robust Method

The robust method is a statistical procedure that is not very sensitive to deviations from the underlying assumptions. In research, outlier data is often obtained which can result in large errors (Bali, Boente, Tyler, & Wang, 2011). The occurrence of cases like this causes the assumptions in the regression analysis to be not fulfilled. So that the interpretation of the model becomes wrong if it continues to apply the Ordinary Least Squares (OLS) method directly (Huber & Ronchetti, 2011).

To overcome outlier data, regression analysis can be done by applying regression coefficient estimation methods that are proven to be robust against outliers data, namely the Alternating Least Square (ALS) development method. This method utilizes a regression approach by iterating the eigenvalues and eigenvectors of an $\boldsymbol{X}$ matrix measuring $m \times n$ (Ren, Li, & Haupt, 2017).

5. Outlier Data

Outlier data appear as a result of recording errors or data processing or regression modeling. Due to the existence of outlier data, it causes a large variety and has the potential to influence the predictive model so that the resulting regression model cannot be relied on because this outlier data will make the estimated regression line drawn disproportionately to the outlier data (Härdle & Simar, 2015).

To overcome this outlier data, if the value is extreme enough, the data is often removed or discarded. In the case of less extreme outlier data values, researchers often hesitate to decide whether to exclude or not.

Outliers data is divided into two parts, namely major outliers and minor outliers. The major outliers data are located in the area outside the $3 \times JAR$ (3 times the distance between quartiles), while the minor outliers data are located in the$1,5 - 3 \times JAR$ (1.5 minuses 3 times the distance between quartiles (Filzmoser & Gregorich, 2020).

C. RESEARCH METHOD

In this study, the authors used data containing minor outliers as presented in Table 1.

Table 1. Kandungan Hara Pupuk Organik Segar (%)

Type of fertilizer	N	P_2O_5	K_2O	CaO	MgO
Cows	2	1.5	2	4	1
Horses	2	1.5	1.5	1.5	1
Goat/Sheeps	2	1.5	3	5	2
Poultries	5	3	1.5	4	1
Wastewater Sludge	2	2	0	2	0.5
Guano	12	12	2.6	11	1

This research was conducted through the following steps (Zhou & Feng, 2017):

1. Determine the initial estimate of the vector $\boldsymbol{u}$.
2. Regressive the initial estimate vector $\boldsymbol{u}$ against each column $j, j = 1, 2, 3,..., p$ with $min\sum \left|\boldsymbol{x}_{ij} - \beta \boldsymbol{u}_{ij}\right|$.
3. Normalize the vector $\boldsymbol{\beta}_j$ with the equation $\boldsymbol{a}_1 = \frac{\beta_j}{\|\beta_j\|}$.
4. Reshape the vector $\boldsymbol{a}_1$ against each row $i, i = 1, 2, ..., n$ matrix $\boldsymbol{X}$ with $min\sum \left|\boldsymbol{x}_{ij} - \beta_i \boldsymbol{a}_{1j}\right|$.
5. Normalize the vector β_i with the equation $\boldsymbol{u}_1 = \frac{\beta_i}{\|\beta_i\|}$.
6. Perform an iterative process to get convergent results.
7. After obtaining the first triple eigen, calculate the error to find the next triple eigen,

$$\varepsilon_i = min\sum\sum \left|\boldsymbol{x}_{ij} - \lambda \boldsymbol{a}_{\boldsymbol{1i}} \boldsymbol{u}_{ij}\right|$$

D. RESULTS AND DISCUSSION

In the case of minor outliers data, the $\boldsymbol{X}$ data matrix measuring 6×5 is used which has been corrected for its mean as follows:

$$\begin{bmatrix} -2.5 & -2.4 & -0.1 & 0.4 & -0.1 \\ -2.5 & -2.4 & -0.6 & -4.1 & -0.1 \\ -2.5 & -2.4 & 2.9 & 1.4 & 0.9 \\ 2.5 & 1.1 & -0.6 & 0.4 & -0.1 \\ -2.5 & -1.9 & -2.1 & -3.6 & -0.6 \\ 7.5 & 8.1 & 0.5 & 5.4 & -0.1 \end{bmatrix}$$

Based on the Singular Value Decomposition, the singular value is obtained

$$\sigma_1 = 14.7; \sigma_2 = 5.6; \sigma_3 = 1.8; \sigma_4 = 1.2; \sigma_5 = 0.04$$

The orthogonal matrix is the eigenvector of $\boldsymbol{X^TX}$ is the matrix $\boldsymbol{A}$ denoted by

$$\begin{bmatrix} 0.6 & -0.3 & 0.2 & 0.7 & -0.1 \\ 0.6 & -0.3 & 0.1 & -0.7 & 0.1 \\ 0.05 & 0.6 & 0.7 & -0.1 & -0.3 \\ 0.5 & 0.7 & -0.6 & 0.0 & 0.0 \\ -0.0 & 0.2 & 0.3 & 0.1 & 0.9 \end{bmatrix}$$

The orthogonal matrix which is the eigenvector of the matrix $\boldsymbol{XX^T}$ is the matrix $\boldsymbol{U}$ denoted by

$$\begin{bmatrix} -0.2 & 0.3 & -0.7 & -0.0 & -0.4 \\ -0.3 & -0.3 & 0.6 & -0.1 & -0.5 \\ -0.2 & 0.7 & 0.3 & -0.1 & 0.3 \\ 0.2 & -0.2 & 0.0 & 0.9 & 0.1 \\ -0.3 & -0.5 & -0.2 & -0.3 & 0.6 \\ 0.8 & -0.1 & 0.0 & -0.3 & -0.1 \end{bmatrix}$$

Obtained matrix $\boldsymbol{X}_{predict} = \boldsymbol{ULA^T}$ as follows:

$$\begin{bmatrix} -2.5 & -2.4 & -0.1 & 0.4 & -0.1 \\ -2.5 & -2.4 & -0.6 & -4.1 & -0.1 \\ -2.5 & -2.4 & 2.9 & 1.4 & 0.9 \\ 2.5 & 1.1 & -0.6 & 0.4 & -0.1 \\ -2.5 & -1.9 & -2.1 & -3.6 & -0.6 \\ 7.5 & 8.1 & 0.5 & 5.4 & -0.1 \end{bmatrix}$$

The steps for decomposing a robust singular value decomposition are:

1. The initial estimate of $\boldsymbol{u}$

$$\boldsymbol{u}_0 = \begin{bmatrix} -0.2 \\ -0.3 \\ -0.2 \\ 0.2 \\ -0.3 \\ 0.8 \end{bmatrix}$$

2. The initial estimate vector regression equation $\boldsymbol{u}$ over all columns in the matrix $\boldsymbol{X}$

$$x_{i1} = 8.9\boldsymbol{u} + \varepsilon_{i1};\ x_{i2} = 9.6\boldsymbol{u} + \varepsilon_{i2};\ x_{i3} = 0.6\boldsymbol{u} + \varepsilon_{i3};\ x_{i4} = 6.5\boldsymbol{u} + \varepsilon_{i4};\ x_{i5} =- 0.1\boldsymbol{u} + \varepsilon_{i5}$$

3. Estimator coefficient of regression

$$\boldsymbol{\beta}_a = \begin{bmatrix} 8{,}9 \\ 9{,}6 \\ 0{,}6 \\ 0{,}2 \\ 6{,}5 \\ -0{,}1 \end{bmatrix}$$

4. Normalize the β_a vector to get the singular value

$$\sigma_a = \sqrt{8{,}9^2 + \ldots + (-0{,}1)^2} = 14{,}7$$

5. Vector $\boldsymbol{a}_1 = \frac{\beta_a}{\sigma_a}$

$$\boldsymbol{a}_1 = \begin{bmatrix} 0{,}6 \\ 0{,}6 \\ 0{,}0 \\ 0{,}4 \\ -0{,}0 \end{bmatrix}$$

6. The vector regression equation $\boldsymbol{a}_1$ against all rows in the matrix $\boldsymbol{X}$

$$x_{1j} =- 3.6\boldsymbol{a}_1 + \varepsilon_{1j}; x_{2j} =- 4.1\boldsymbol{a}_1 + \varepsilon_{2j};\ x_{3j} =- 3.7\boldsymbol{a}_1 + \varepsilon_{3j};\ x_{4j} = 1.6\boldsymbol{a}_1 + \varepsilon_{4j};\ x_{5j} =- 4.0\boldsymbol{a}_1 + \varepsilon_{5j};\ x_{6j} = 12.3\boldsymbol{a}_1 + \varepsilon_{6j}$$

7. Estimator coefficient of regression

$$\boldsymbol{\beta}_u = \begin{bmatrix} -3.7 \\ -4.1 \\ -3.7 \\ 1.6 \\ -4.1 \end{bmatrix}$$

8. Normalize the β_u vector to get the singular value

$$\sigma_u = \sqrt{(-3.7)^2 + \ldots + (-4.1)^2} = 14.6$$

9. Vector $\boldsymbol{u}_1 = \frac{\beta_u}{\sigma_u}$

$$\boldsymbol{u}_1 = \begin{bmatrix} -0.2 \\ -0.3 \\ -0.2 \\ 0.1 \\ -0.3 \\ 0.8 \end{bmatrix}$$

This process is carried out by iteration to obtain convergent results. In the first triple eigen, this is done in three iterations so that the results are convergent. Based on the results of the analysis, it was obtained the fifth triple eigen which was carried out in four iterations to obtain convergent results. Thus the results of the robust singular value decomposition are:

1. Diagonal matrix $\boldsymbol{L}$

$$\boldsymbol{L} = \begin{bmatrix} 14.6 & 0 & 0 & 0 & 0 \\ 0 & 5.3 & 0 & 0 & 0 \\ 0 & 0 & 2.7 & 0 & 0 \\ 0 & 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 0 & 0.8 \end{bmatrix}$$

2. Orthogonal matrix $\boldsymbol{A}$

$$\boldsymbol{A} = \begin{bmatrix} 0.6 & -0.05 & 0.1 & 0.9 & 0.0 \\ 0.6 & -0.0 & 0.0 & -0.0 & 0.9 \\ 0.04 & 0.7 & -0.0 & -0.2 & -0.04 \\ 0.4 & 0.7 & -0.9 & 0.0 & -0.0 \\ -0.0 & 0.2 & 0.06 & 0.01 & -0.4 \end{bmatrix}$$

3. Orthogonal matrix $\boldsymbol{U}$

$$\boldsymbol{U} = \begin{bmatrix} -0.2 & 0.01 & -0.7 & 0.0 & -0.0 \\ -0.3 & -0.1 & 0.6 & -0.18 & 0.3 \\ -0.2 & 0.8 & 0.0 & -0.0 & 0.0 \\ 0.1 & -0.1 & 0.0 & 0.9 & 0.0 \\ -0.3 & -0.5 & -0.05 & -0.09 & 0.9 \\ 0.8 & -0.0 & 0.0 & -0.0 & -0.0 \end{bmatrix}$$

Then the description matrix through the robust singular value decomposition method is multiplied back to $\boldsymbol{ULA^T}$ to get $\boldsymbol{X}_{predict}$.

$$\boldsymbol{X} = \begin{bmatrix} -2.5 & -2.4 & -0.1 & 0.4 & -0.08 \\ -2.5 & -2.4 & -0.6 & -4.1 & -0.08 \\ -2.5 & -2.4 & 2.9 & 1.4 & 0.9 \\ 2.5 & 1.1 & -0.6 & 0.4 & -0.08 \\ -2.5 & -1.9 & -2.1 & -3.6 & -0.6 \\ 7.5 & 8.1 & 0.5 & 5.4 & -0.08 \end{bmatrix}$$

In the minor outlier data, the $\boldsymbol{X}_{predict}$ obtained through the robust singular value decomposition method is the same as the actual $\boldsymbol{X}$.

E. CONCLUSION AND SUGGESTION

Based on the results of the analysis using the singular value decomposition method and the robust singular value decomposition method, the completion of data clusters containing minor outliers in the sample data on the nutrient content of organic fertilizers using the robust singular value decomposition method, has better results than using the singular value decomposition method. The suggestion for further research is to examine the RSVD method on major outlier data.

ACKNOWLEDGEMENTS

Researchers would like to thank the STMIK Pringsewu for giving permission to collaborate in the field of research and provide moral support, so that this research can be carried out well.

REFERENCES

Aa, N. . Van der, Morsche, H. G. Ter, & Mattheij, R. R. M. (2007). Ela Computation of Eigenvalue and Eigenvector. *Electronic Journal Of Linear Algebra*, *16*(1/2), 300–314.

Anton, H. (1987). *Aljabar Linear Elementer* (Jakarta). Erlangga.

Bali, J. L., Boente, G., Tyler, D. E., & Wang, J. L. (2011). Robust Functional Principal Components: A Projection-Pursuit Approach. *The Annals of Statistics*, *39*(6), 2852–2882.

Bretscher, O. (1997). *Linear Algebra with Applications*. New York: Prentice-Hall Inc.

Filzmoser, P., & Gregorich, M. (2020). Multivariate Outlier Detection in Applied Data Analysis: Global, Local, Compositional and Cellwise Outliers. *Mathematical Geosciences*, *12*(2), 1–18.

Härdle, W. K., & Simar, L. (2015). *Applied Multivariate Statistical Analysis* (4th ed.). Berlin: Springer.

Huber, P. J., & Ronchetti, E. M. (2011). *Robust Statistics* (2nd ed.). New Jersey: John Wiley & Sons.

Liu, L., Hawkins, D. M., Gosh, S., & Young, S. S. (2003). Robust Singular Value Decomposition Analysis of Microarray Data. *Proceedings of the National Academy of Sciences of the USA*, *100*, 167–172.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., Cohen, K. L., … Cohen, K. L. (1999). Robust Principal Component Analysis for Functional Data. *Test*, *8*(1), 1–73.

Neter, J., Wasserman, W., and Kutner, M. H. (1990). *Applied Linear Statistical Model*. New York: Richard D Irwin Inc.

Ren, J., Li, X., & Haupt, J. (2017). Robust PCA via Tensor Outlier Pursuit. *Conference Record - Asilomar Conference on Signals, Systems and Computers*, 1744–1749.

Valverde-albacete, & J, F. (2020). The Singular Value Decomposition over Completed Idempotent Semifields. *Mathematics*, *8*(9), 1–39.

Zhang, L., Marron, J.S., Shen, H., and Z. Z. (2007). Singular Value Decomposition and Its Visualization. *Journal of Computational Graphical Statistics*, *16*, 833–854.

Zhang, L., Shen, H., Huang, J. Z. (2013).). Robust Regularized Singular Value Decomposition with Application to Mortality Data. *The Annals of Applied Statistics*, *7*(3), 1–23.

Zhou, P., & Feng, J. (2017). Outlier Robust Tensor PCA. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2263–2271.