



# **iJRASET**

International Journal For Research in  
Applied Science and Engineering Technology



---

# **INTERNATIONAL JOURNAL FOR RESEARCH**

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume: 6      Issue: IV      Month of publication: April 2018**

**DOI: <http://doi.org/10.22214/ijraset.2018.4415>**

**[www.ijraset.com](http://www.ijraset.com)**

**Call:  08813907089**

**E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)**

# Subject Grouping Using Mahalanobis Distance and PCA

Melda Putri Boni<sup>1</sup>, Sutarman, Dan Saib Suwilo<sup>2</sup>

<sup>1</sup>Himpunan Mahasiswa Matematika

**Abstract.** *In multivariate analysis PCA can be used to reduce data so that data that initially has many variables will produce a few variables making it easier to do the grouping, and with the PCA then the multicollinearity data can be overcome. The use of PCA is the Robust PCA (ROBPCA) so that data is not easily influenced by outliers. With PCA reductions, clustering with cluster analysis using Mahalanobis Distance similarity will result in more optimal grouping, since the outliers have been overcome with Robust using Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE).*

**Keywords:** *Principal Component Analysis, ROBPCA, Mahalanobis Distance, MCD, MVE*

## I. INTRODUCTION

Multivariate analysis is a method of processing variables in large numbers, where the goal is to find the influence of these variables on an object simultaneously or simultaneously. The classification of multivariate analysis, namely dependensi method, and interpedensi method. Some examples of interferensi analysis, among others Factor Analysis, Multidimensional Scale Analysis, Cluster Analysis, and Principal Component Analysis. Cluster analysis was first used by Tyron in 1939. Cluster analyst is one of the analytical methods used in multivariate, in Cluster Analysis there is no independent variable and independent variables. The objects within each group tend to resemble each other and differ greatly from the objects of the other clusters. There are various sizes that can state that certain objects have similarities, correlations, distances, and associations (Yang, 2004). This study uses distance similarity measure, ie Mahalanobis Distance. One example of the use of Cluster Analysis is at the level of air pollution examined by Rachmatin in 2014, in that study researchers using the similarity measure is Euclidian Distance. The researcher suggested to use other similarity measure of distance, so that the clustering result better, because if only using Euclidian Distance less effective grouping. Euclidian Distance has advantages can be used to reflect the inequalities of two patterns, and can be used to evaluate the proximity of two or three dimensional data objects. Euclidian Distance is very sensitive to the size of the sample and the magnitude of the distribution of variance, and this use is not effective if there are correlations between variables so that Principal Component Analysis can be used to eliminate correlations between variables (Wichern, 2002: 670). Euclidian Distance can not be used in the original data, because it can not anticipate scale changes. At different scales the distance obtained will be very different as well so that the results of the Cluster Analysis to categorize the subject will also be different. Based on the advantages and disadvantages, the authors are interested to use the similarity measure of distance is another, namely Mahalanobis Distance. Mahalanobis Distance is a generalization of standardized Euclidian Distance, in Mahalanobis Distance considering the variability elements (variance), which involves the matrix of variancovarians in calculating the distance, when the variable is correlated both positive and negative, Mahalanobis Distance measurement becomes the most appropriate use because adjust the intercorrelation. In Cluster Analysis it is difficult to categorize if there is an outlier and multicollinearity, so the need for Principal Component Analysis (PCA), because PCA aims to reduce existing variables become less without losing information in original or initial data. With the PCA, the initial variable  $n$  will be reduced to  $k$  new variables, with the number of  $k < n$  variables. With  $k$  variables already represented for  $n$  variables, the resultant variable is called Principal Component, so grouping will be easier and optimal. Outliers that occur can be analyzed with Robust, this method is an important tool for analyzing data that is influenced by outliers so as to produce Robust model or resistance to outlier. A resistance estimate is relatively unaffected by large changes in small pieces of data or small changes in large parts of the data. Based on the problems, and the weaknesses of the various similarity measures of distance and also the disadvantages of the Cluster Analysis described above, the authors are interested in raising the title of the study "Grouping Subjects Using Mahalanobis Distance and Principal Component Analysis".

## II. LITERATURE AND REVIEW

Multivariate analysis is a treating statistical analysis technique a group of criterion variables that are correlated as one system, with taking into account the correlation between those variables ( Richard and Dean, 2007 ). In the cluster analysis, before clustering the object should be note the size of its similarity (Wu and Yang, 2004). The equation form of Euclidian Distance is as following:

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

where:  $j$  = Euclidian Distance of the  $i$  data and the  $j$ th data object.  $m$  = Many variables or parameters used.  $x_{ik}$  = Object of  $i$ th data on  $k$ -th variable  $x_{jk}$  = the data object to  $j$  on the  $k$ .

Mahalanobis Distance is a generalization of the Euclidian Distance distandardization. Distance between individuals  $S_i$  and individual  $S_j$  are expressed by:

$$d_{ij}^2 = (x_i - x_j)T \Sigma^{-1} (x_i - x_j)$$

Principal Component Analysis is basically aimed at to simplify the observed variables by reducing their dimensions ( Johnson dan Winchern, 2007). Robust regression is a tool that can be used to analyze data which contain an outlier and provide resistant results to the presence outlier (Turkan, Cetin, and Toktamis, 2012)

### III. RESULTS AND DISCUSSION

#### A. Application of ROBPCA with MCD and MVE

Implementation on ROBPCA using MCD and MVE, use pollution data, where the data consists of 60 observations by number variable 16, this data is solved by using program R. The data, will be reduced by using PCA, here is:

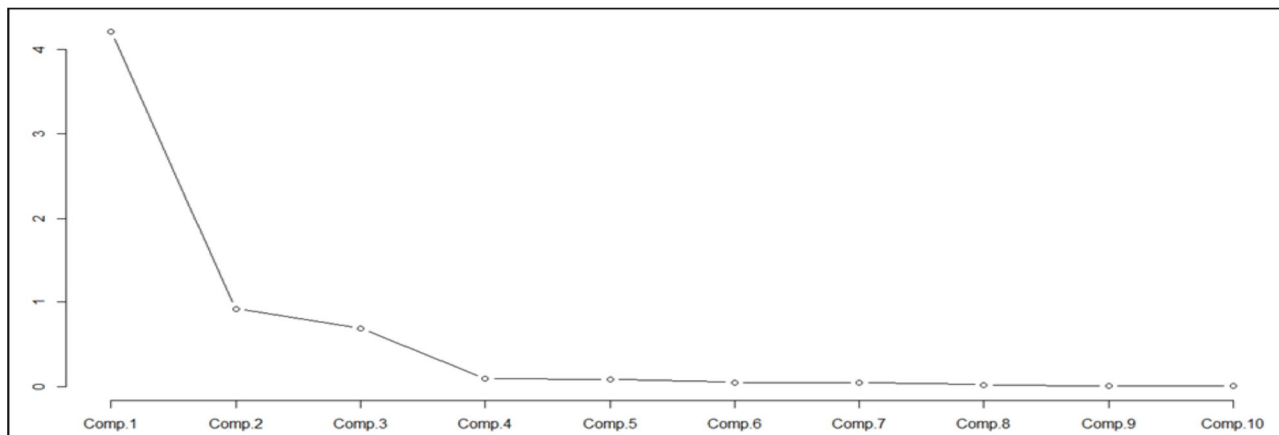
```

Importance of components:
      Comp.1      Comp.2      Comp.3      Comp.4      Comp.5
Standard deviation  2.0515878  0.9601621  0.8303066  0.30864741  0.29645778
Proportion of Variance  0.6839359  0.1498043  0.1120243  0.01547963  0.01428107
Cumulative Proportion  0.6839359  0.8337402  0.9457645  0.96124412  0.97552519
      Comp.6      Comp.7      Comp.8      Comp.9
Standard deviation  0.225252923  0.215384111  0.150123197  0.108216653
Proportion of Variance  0.008244723  0.007538111  0.003662105  0.001902932
Cumulative Proportion  0.983769914  0.991308025  0.994970130  0.996873063
      Comp.10     Comp.11     Comp.12     Comp.13
Standard deviation  0.093809126  0.073611950  0.0425411310  0.0347306836
Proportion of Variance  0.001429965  0.000880505  0.0002940717  0.0001960026
Cumulative Proportion  0.998303028  0.999183533  0.9994776044  0.9996736070
      Comp.14     Comp.15     Comp.16
Standard deviation  0.0316706462  0.0263457233  1.765020e-02
Proportion of Variance  0.0001629855  0.0001127861  5.062143e-05
Cumulative Proportion  0.9998365925  0.9999493786  1.000000e+00
    
```

#### B. Image of PCA Results

Visible from the proportion of cumulative variance The first component can be explored 0.68 total variance and when added the second component, and third to 0.94. This means that if we only take 3 components only the first component up to the 3rd component is sufficient. The

following is the scree-plot of the PCA, because in the presence of these scots can be known that which components are sufficien is the scree-plot of the PCA, because in the presence of these scots can be known that which components are sufficiently representative of those 16 variables:





C. Image Scree-Plot

From the scree plot it is seen that the curve starts to ramp up at the point of comp 4 that with three components alone is sufficient to represent the 16 variable. Three components are formed which represent 16 variables, had it going will be shown the results of ordinary covarian, using MCD and MVE.

Tabel 1. Mahalanobis Distance

5.300	22.213	7.6	8.341	9.718.04	151.770.7	132.642.5	6.589.7	7.250.5	9.505.34	52.963.167
		4		6	54	46	1	4	2	
8.362	9.033	6.7	13.209	2.641	268.798.4	28.084.36	10.497.	26.154.	6.324.53	10.329.0
		7	.		19	7	8	1	8	16
49.71	8	9.9	9.759.	14.084.7	50.567.88	112.052.1	207.15	11.011.	8.913.73	6.707.934
	5.400	7	4	5	0	97	4.	2	47.891.5	9
115.8	4	8.9	120.89	144.742.	7.187.572	22.530.86	10.254.	6.279.2	749.401.	3.751.258.2
	10.033	9	5	0		9	0	0	11.459.7	91
488.9	165.14	5.6	9.368.	7.286.96	9.426.485	73.560.43	11.305.	8.556.7	13.480.7	8.259.493
	7	4	3	4		8	5	0	10.073.1	06

Tabel 2. Mahalanobis Distance With MCD

2.05	3.244.0		3.191.	5.042.6	11.162.	7.758.	3.032.7	3.071.59	1.898.7	3.554.2	4.249.	2.841.
4.7	19	3.537.514	083	18	981	948	34	3	73	41	070	750
1.21	1.781.1		3.520.	28.487.	1.780.1	5.100.	5.213.7	3.771.21	3.533.0	4.577.4	4.929.	3.824.
4.9	62	3.222.804	039	268	35	407	84	6	34	31	996	733
3.39	2.752.0	1.296.620	6.345.	9.738.4	18.672.	4.689.	4.796.2	3.736.99	3.127.0	8.136.1	2.172.	3.826.
9.9	85	.456	082	34	995	863	74	9	93	69	625	260
10.5	11.672.		4.648.	4.883.9	2.989.9	3.761.	67.525.	361.503.	40.432.	14.168.	3.513.	4.254.
49	796	2.814.529	288	90	06	899	068	426	000	981	977	572
3.13	3.293.0		3.856.	3.736.1	3.716.8	4.468.	1.939.1					
2.3	11	4.332.031	629	37	49	233	74					

Tabel 3. Mahalanobis Distance With MVE

[1]	5.004.68	66.38801	7.426.60	6.709.98	14.506.6	3.819.78	39.46642	4.347.47	6.213.18	3.454.1	2.934.4
	1	53.38032	9	8	33	3	40.49256	7	4	14	03
[1	2.038.36	27.47848	5.927.69	22.574.2	3.169.07	10.022.5	107.1510	5.056.38	5.488.85	3.400.7	4.193.7
4]	3	37.45691	1	82	3	73	67.35472	9	7	69	99
[2	3.846.80	14.41322	11.780.7	6.663.73	3.393.10	9.163.29	76.41869	4.235.13	12.244.9	2.760.1	7.550.7
7]	2	27428.19	71	5	7	8	74.45314	0	92	90	98
[4	14.070.4	123.1625	7.992.87	7.991.82	4.343.72	1.800.08	891.1290	32.650.2	11.092.8	9.540.2	5.777.3
0]	21	85.91509	0	2	6	0	7	29	99	80	27
[5	5.697.89	29.84790	5.199.92	3.499.02	12.520.3	7.557.24	5.266.87				
3]	0	-18.35844	5	4	20	7	2				

The robust data, it can be concluded that the results from clustering will get better. With MCD and MVE the data is no longer influenced by the outliers, and the distance of the mahalanobis is different according to the covariance of MCD and Covarian MVE. The better distinguishing mahalanobis the MCD and MVE than the classical ones. Because no longer influenced outlier.



#### IV. CONCLUSION

With robust PCA, which uses MCD and MVE, grouping of subjects becomes easier, more efficient and optimal. Grouping is no longer changing because it is not influenced by outliers. Mahalanobis distance becomes more efficient with MCD and MVE.

#### REFERENCES

- [1] Aelst, S. V dan Rousseeuw, P. (2009). Minimum Volume Ellipsoid. *Wires Computational Statistics* volume 1, 71-82. Everitt, B. S. (1980). *Cluster Analysis for Application*. Second Edition. Heineman Educational Books Ltd, London.
- [2] Hubert, M., Rousseeuw, P. J., dan Vanden Branden, K. (2005), ROBPCA: A new Approach to Robust Principal Component Analysis, *Technometrics*, 47: 64 - 78.
- [3] Hubert, M. dan Debruyne, M. (2009). Minimum Covariance Determinant. *Wires Computational Statistics* 2010, 36-43.
- [4] Johnson, R. A. dan Wichin, D. W. (2002). *Applied Multivariate Statistical Analysis*. Edisi 5. New York: Prentice Hall.
- [5] Johnson, R. A. dan Wichin, D. W. (2007). *Applied Multivariate Statistical Analysis*. New York: Prentice Hall.
- [6] Maronna, R. A. (1976), Robust Estimator of Multivariate Location and Scatter, *Ann. Statist.*, 4: 51-67.
- [7] McLachlan, G. J. Kay, Shu. dan Bean, Richard. (2006). Robust Cluster Analysis Via Mixture Models. *Australian Journal Of Statistics*. Volume 35, Issue 2 and 3, 157-174
- [8] Rousseeuw, P.J. (1999). Fast Algorithm for the Minimum Covariance Determinant Estimator. *American Statistical Association and the American Society for Quality*, Volume 41, Issue 3, 212-223
- [9] Yang, M. S. dan Wu, K. L. (2004). A Similarity-Based Robust Clustering Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Volume 26, Issue 4, 434-435.



10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)