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

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Subject Grouping Using Mahalanobis Distance and PCA

Melda Putri Boni[1], Sutarman, Dan Saib Suwilo[2]
[1]*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 multicolloniarity 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 Analysist, 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 simultaneously. The classification of multivariate analysis, namely depedensi 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 Eucledian Distance. The researcher suggested to use other similarity measure of distance, so that the clustering result better, because if only using Eucledian Distance less effective grouping. Eucledian 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. Eucledian 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). Eucledian 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 Eucledian 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 multicolloniarity, 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 Eucledian Distance is as following:

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

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

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

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

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*

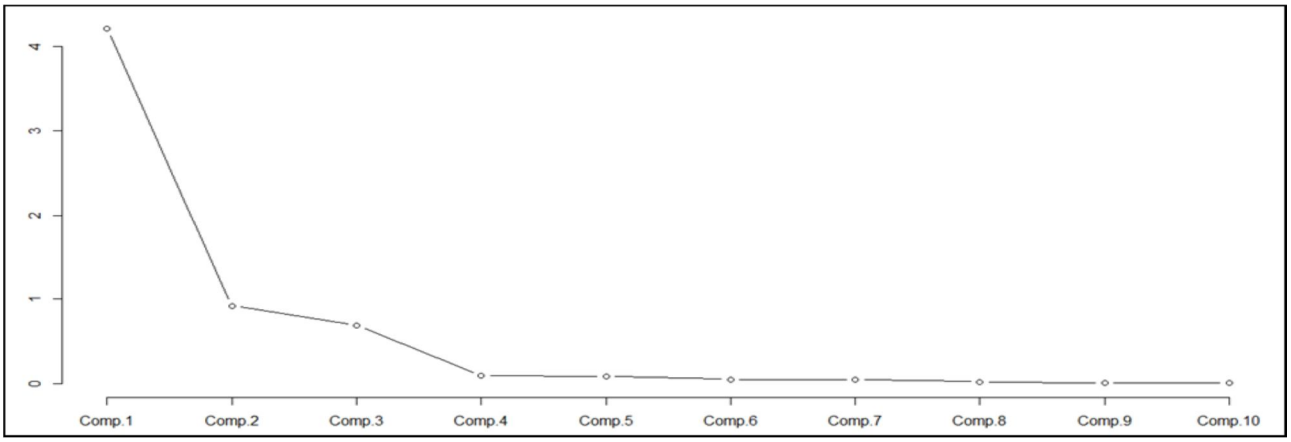
Implementation on ROBPCA using MCD and MVE, use pollution data, where the data consists of 60 observations by number variable 16, this data is solved by using program R. The data, will be reduced by using PCA, here is:

```
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	7.64	8.341	9.718.046	151.770.754	132.642.546	6.589.71	7.250.54	2.678.19	9.505.342	52.963.167
8.362	9.033	6.77	13.209.	2.641	268.798.419	28.084.367	10.497.8	26.154.1	6.324.538	10.329.016	9.706.986
49.718	5.400	9.97	9.759.4	14.084.75	50.567.880	112.052.197	207.154.	11.011.2	47.891.5	8.913.739	6.707.934
115.84	10.033	8.99	120.895	144.742.0	7.187.572	22.530.869	10.254.0	6.279.20	11.459.7	749.401.91	3.751.258.201
488.92	165.147	5.64	9.368.3	7.286.964	9.426.485	73.560.438	11.305.5	8.556.70	10.073.1	13.480.706	8.259.493

Tabel 2. Mahalanobis Distance With MCD

2.054.7	3.244.019	3.537.514	3.191.083	5.042.618	11.162.981	7.758.948	3.032.734	3.071.593	1.898.773	3.554.241	4.249.070	2.841.750
1.214.9	1.781.162	3.222.804	3.520.039	28.487.268	1.780.135	5.100.407	5.213.784	3.771.216	3.533.034	4.577.431	4.929.996	3.824.733
3.399.9	2.752.085	1.296.620.456	6.345.082	9.738.434	18.672.995	4.689.863	4.796.274	3.736.999	3.127.093	8.136.169	2.172.625	3.826.260
10.549	11.672.796	2.814.529	4.648.288	4.883.990	2.989.906	3.761.899	67.525.068	361.503.426	40.432.000	14.168.981	3.513.977	4.254.572
3.132.3	3.293.011	4.332.031	3.856.629	3.736.137	3.716.849	4.468.233	1.939.174					

Tabel 3. Mahalanobis Distance With MVE

[1]	5.004.681	66.38801 53.38032	7.426.609	6.709.988	14.506.633	3.819.783	39.46642 40.49256	4.347.477	6.213.184	3.454.114	2.934.403
[14]	2.038.363	27.47848 37.45691	5.927.691	22.574.282	3.169.073	10.022.573	107.15104 67.35472	5.056.389	5.488.857	3.400.769	4.193.799
[27]	3.846.802	14.41322 27428.19840	11.780.771	-6.663.735	-3.393.107	9.163.298	76.41869 74.45314	4.235.130	12.244.992	2.760.190	7.550.798
[40]	14.070.421	123.16253 85.91509	7.992.870	7.991.822	4.343.726	1.800.080	891.12907 8784.65429	32.650.229	11.092.899	9.540.280	5.777.327
[53]	5.697.890	29.84790 -18.35844	5.199.925	3.499.024	12.520.320	7.557.247	5.266.872				

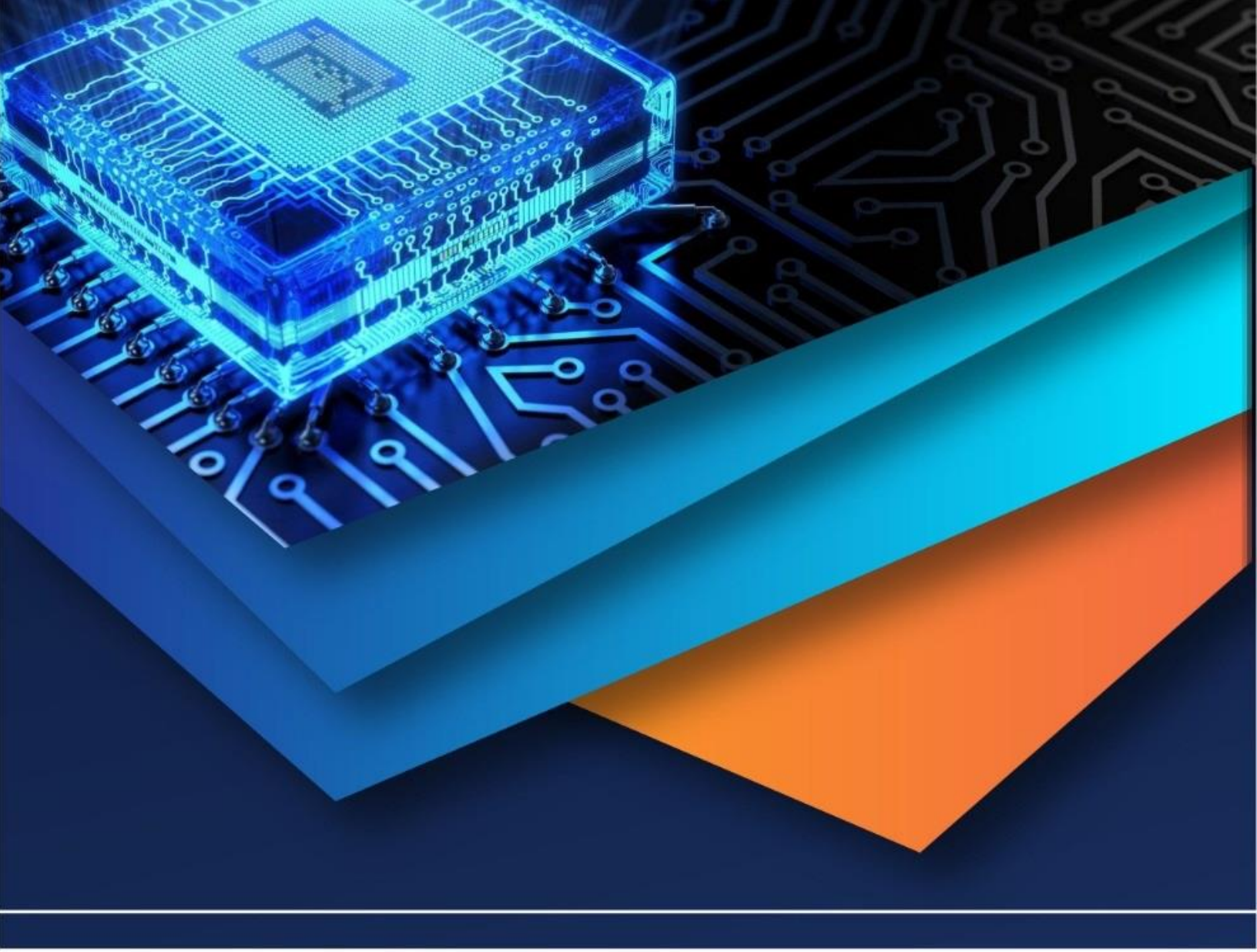
The robust data, it can be concluded that the results from clustering will get better. With MCD and MVE the data is no longer influenced by the outliers, and the distance of the mahalanobis is different according to the covariance of MCD and Covarian MVE. The better distinguishing mahalanobis the MCD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

IV. CONCLUTION

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 Aplication. 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 Componen 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 Wichen, D. W. (2002). Aplied Multivariate Statistical Analysis. Edisisi 5. New york: Prentice Hall.
[5] Johnson, R. A. dan Wichen, D. W. (2007). Aplied 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. Australia 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 Intellegence 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