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 V, May 2018- Available at www.ijraset.com

Object Classification with Classical Linear Discriminant Analysis and Robust Linear Discriminant Analysis

Justin Eduardo Simarmata[1], Sutarman[2], Mardiningsih[3]
[1, 2, 3]*Faculty of Mathematics and Science, University of Sumatera Utara*

Abstract: Discriminant analysis is one of multivariate analysis with dependency method. Discriminant analysis is a multivariate analysis that aims to classify observations based on several independent variables that are non-categorical and categorical dependent variables. Discriminant analysis requires the assumption of the normal multivariate distribution and the homogeneity of the variance-covariance matrix. Classical linear discriminant analysis and linear robust discriminant analysis can be applied to classify objects. The classification is based on 10 indicators of district/city poverty level in North Sumatra province, 10 indicators are as independent variable and low poverty classification and high poverty classification as dependent variable. The linear robust discriminant model classifies the object more precisely than the classic linear discriminant model. This can be seen from the total proportion of classification mistakes of 6%, less than the total proportion of classical linear discriminant classifier error classification of 21.2%. This is due to the large number of outlays in the district/city poverty data in North Sumatra.
Keywords: multivariate analysis, the classical linear discriminant analysis, robust

I. INTRODUCTION

Multivariate analysis is a statistical analysis that is imposed on the data that consists of many variables and intercorrelated variables. Multivariate data not only consists of one variable, but may consist of more than one variable. Multivariate analysis is a statistical technique used to understand the data structures in high dimensions. The variables are related to one another. There in lies the difference between multivariable and multivariate analyzes. Multivariate inevitably involves a multivariable but not vice versa. Multivariable mutually correlated which is said multivariate.

Discriminant analysis is a multivariate analysis that aims to classify observations based on several independent variables that are non-categorical and categorical dependent variables. Discriminant analysis requires the assumption of multivariate normal distribution and homogeneity of variance-covariance matrix. In the application of discriminant analysis to consider the outliers in the data. Therefore, the average vector and variance-covariance matrices are predicted by the robust minimum covariance determinant (MCD) method of the outliers.

Seeing the many methods of grouping objects that exist to encourage researchers to compare the methods with each other to determine the most accurate method of classifying an object. The model will be established in this study is the classical linear discriminant analysis and linear discriminant analysis robust. The second analysis was chosen as the object classification of the most widely used both theoretically and practically. The analysis is also a kind of analysis of the most widely developed by many researchers because it has a level of accuracy and precision in classifying an object that is higher than any other method types.

This stuy aims to classify the data rate of poverty districts/cities in North Sumatra by linear discriminant analysis classical and discriminant analysis robust, obtain the classification of objects into a group with the method of linear discriminant analysis classical and robust and compare results of the classification method of linear discriminant analysis classic with robust discriminant analysis to obtain the best results based on wrong classification of the minimum.

II. DISCRIMINANT ANALYSIS

Multivariate analysis is an analysis that involves many variables or double variables. Discriminant analysis is one of the multivariate analysis. Discriminant analysis is a statistical technique for classifying individuals into groups that are independent and firmly based cluster of independent variables. Discriminant analysis aims to understand different groups and predict the probability that an object of research will go a certain group members. The purpose of discriminant analysis is to create a linear discriminant function or a combination of predictors or independent variables that can discriminate or distinguish the dependent variable category or group, that is able to distinguish an object enter the group in which categories.

A. *Multivariate Normality Test*

To test the normality of multiple variables is to find the value of squared distance for each observation that is:

$$d_i^2 = (x_i - \tilde{x})^T S^{-1}(x_i - \tilde{x})$$

B. *The Variance Covariance Matrix Similarity Test*

To test the variance covariance matrix similarity group I (S_1) and group II (S_2) used a hypothesis:

$H_0 : S_1 = S_2$, the variance matrix of group covariance is relatively the same

$H_1 : S_1 \neq S_2$, the matrix of covariance variance of the group is significantly different

accept H_0, which means the matrix of covariance variance is the same if:

$$X^2_{count} \leq X^2_{\left(\propto.\left[\frac{1}{2}(k-1)p(p+1)\right]\right)}$$

with:

$$X^2_{count} = -2(I - C_i)\left[\frac{1}{2}\sum_{i=1}^{k} V_i \ln|S_i| - \frac{1}{2}\ln|S|\sum_{i=1}^{k} V_i\right]$$

with:

$$V_i = n_i - 1$$

$$S = \frac{\sum_{i=1}^{k} V_i S_i}{\sum_{i=1}^{k} V_i}$$

$$C_1 = \left[\sum_{i=1}^{k}\frac{1}{V_i} - \frac{1}{\sum_{i=1}^{k} V_i}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right]$$

C. *Variance Difference Variance Testing*

The test statistic used to test the average between groups is statistic F with the hypothesis:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, means the mean between groups is the same (no difference)

$H_1 : \mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$, means there is an average difference between groups

α = Level of Significance

Critical areas : rejected H_0, jika $F_{count} > F_{table}$

$F_{table} = F_{\propto(db1;db2)}$

$db_1 = k - 1$

$db_2 = (n - k) = (n_1 - 1) + (n_2 - 1)$

D. *Outlier Detection*

Outlier is an observation that is far (extreme) from other observations. An observation x_i is detected as an outlier if its mahalanobis distance is as follows:

$$d^2_{MD} = (x_i - \tilde{x})^T S^{-1}(x_i - \tilde{x}) > x^2_{p.(\propto)}$$

E. *Function of Discriminant Analysis*

The discriminant function is a linear combination of variables belonging to groups to be classified. In the *i*-nth observed data (i = 1, 2, ... , *n*) consisting of variables are $X_1, X_2, \ldots, X_j$. The observational data can be presented in the following matrix form.

Table 1. Observational Data Matrix

Variables	X_1	X_2	⋯	X_j
Observation data	X_{11}	X_{12}	⋯	X_{1j}
	X_{21}	X_{22}	⋯	X_{2j}
	⋮	⋮		⋮
	Xn_1	Xn_2	⋯	X_{nj}

For the calculated X_j variable is the variance, given the symbol S_{jj} with the formula:

$$S_{jj} = \frac{n\sum_{n=1}^{n} x_{nj}^2 - \left(\sum_{n=1}^{n} x_{nj}\right)^2}{n\,(n-1)}$$

For the variables X_i and X_j where $i \neq j$ there is covariance, given the symbol S_{ij} which can be calculated by the following formula:

$$S_{ij} = \frac{n\sum_{n=1}^{n} x_{ni}x_{nj} - \left(\sum_{n=1}^{n} x_{ni}\right)\left(\sum_{n=1}^{n} x_{nj}\right)}{n\,(n-1)}$$

Variance and covariance are arranged in a matrix called the variance-covariance matrix with the symbol S:

$$S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1j} \\ S_{21} & S_{22} & \cdots & S_{2j} \\ \vdots & \vdots & \cdots & \vdots \\ S_{j1} & S_{j2} & \cdots & S_{jj} \end{bmatrix}$$

The variables in each group can be written in column vector form as follows.

$$X_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1j} \end{bmatrix} \text{ and } X_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2j} \end{bmatrix}$$

X_{1j} declares the variable X to j in group to 1

X_{2j} declares the X variable to j in group to 2

From each n_1-sized group of the 1st and n_2-sized groups of the 2nd group. Observation data will be in the form of a matrix that looks like the following.

Table 2. Observation Data Matrix from Group I

Variables	X_{11}	X_{12}	$\cdots$	X_{1j}
Observation data	X_{11}	X_{12}	$\cdots$	X_{1j}
	X_{21}	X_{22}	$\cdots$	X_{2j}
	$\vdots$	$\vdots$		$\vdots$
	Xn_1	Xn_2	$\cdots$	X_{nj}
Average	$\bar{X}_{11}$	$\bar{X}_{12}$	$\cdots$	$\bar{X}_{1j}$

Table 3. Observation Data Matrix from Group I

Variables	X_{21}	X_{22}	$\cdots$	X_{2j}
Observation data	X_{211}	X_{211}	$\cdots$	X_{2j1}
	X_{212}	X_{222}	$\cdots$	X_{2j2}
	$\vdots$	$\vdots$		$\vdots$
	$X_{21}n_2$	$X_{22}n_2$	$\cdots$	$X_{2j}\ n_2$
Average	$\bar{X}_{11}$	$\bar{X}_{12}$	$\cdots$	$\bar{X}_{2j}$

The results of this observation will produce the mean for each variable formed in vector form can be written:

$$\bar{X}_1 = \begin{bmatrix} \bar{X}_{11} \\ \bar{X}_{12} \\ \vdots \\ \bar{X}_{1j} \end{bmatrix} \text{ and } \bar{X}_2 = \begin{bmatrix} \bar{X}_{21} \\ \bar{X}_{22} \\ \vdots \\ \bar{X}_{2j} \end{bmatrix}$$

where:

X_{1jn1} declares variable X to j in group to 1 of size n_1

X_{2jn2} declares variable X to j in group 2 that is n_2 sized

$\bar{X}_{1j}$ states the average of variables to j in group 1

$\bar{X}_{2j}$ states the average of variables to j in group 2

The variance of covariance is arranged in matrix S_1 and S_2, that is:

$$S_1 = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1j} \\ S_{21} & S_{22} & \cdots & S_{2j} \\ \vdots & \vdots & \cdots & \vdots \\ S_{j1} & S_{j2} & \cdots & S_{jj} \end{bmatrix} \quad \text{and} \quad S_2 = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1j} \\ S_{21} & S_{22} & \cdots & S_{2j} \\ \vdots & \vdots & \cdots & \vdots \\ S_{j1} & S_{j2} & \cdots & S_{jj} \end{bmatrix}$$

where : S_1 = matrix of covariance variance of group 1
S_2 = matrix of covariance variance of group 2

Both matrices of this variance-covariance matrix can be calculated by the combined variance-covariance matrix, given the symbol S with the formula:

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

This combined variance-covariance matrix has an inverse, S^{-1}.

Given the mean vectors $\bar{X}_1$ and $\bar{X}_2$ and also the combined variance-covariance matrix S, the discriminant function formula is obtained:

$$Y = (\bar{X}_1 - \bar{X}_2)\, S^{-1}.X$$

F. Robust Method

To overcome the existence of the pencil then used Minimum Covariance Determinant (MCD) estimator for μ and Σ which allegedly by $\bar{x}$ and S is

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad \text{and} \quad S = \frac{\sum_{i=1}^{n} w_i(x_i-\bar{x})((x_i-\bar{x}))^t}{\sum_{i=1}^{n} w_i - 1}$$

thus, the discriminant score equation for RLDA becomes:

$$d_k^l(x) = \bar{x}_k^t S^{-1} x - \frac{1}{2}\bar{x}_k^t S^{-1}\bar{x}_k + \ln(p_k).k = 1.2.\dots.g$$

and the discriminant score equation for RQDA becomes:

$$d_k^Q(x) = -\frac{1}{2}\ln|S_k| - \frac{1}{2}(x - \bar{x}_k)^t S_k^{-1}(x - \bar{x}_k) + \ln(p_k).k = 1.2.\dots.g$$

an observation x will be included in a group of k if the discriminant score:

$$d_x(x) = \text{maximum from } d_1(x).d_2(x) \dots d_g(x)$$

G. Apparent Error Rate (APER)

APER value can be calculated by:

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

where:

n_{1M} = the number of observations of group 1 were incorrectly classified as group 2

n_{2M} = the number of observations from group 2 were incorrectly classified as group 1

H. The accuracy Grouping Discriminant Function

Table 4. Evaluation of Discriminant Functions

Initial Grouping	Grouping By Discriminant Functions		Total
	Group I	Group II	
Group I	n_{11}	n_{12}	$n_{1,}$
Group II	n_{21}	n_{22}	$n_{2,}$
Total	$n_{,1}$	$n_{,2}$	N

III. RESULTS AND DISCUSSION

The data used is secondary data. The secondary data is obtained from the national economic census data of 2016. Secondary data used is the percentage of population per-districts/cities based on indicators of poverty in North Sumatera used as a variable X to measure the poverty level Y. Data obtained at the district/city level is in North Sumatra consisting of 33 districts/cities.

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 V, May 2018- Available at www.ijraset.com

District/Municipality Poverty Rate Data in North Sumatra:

Table 5. Poverty Rate of regencies/cities data in North Sumatra

Number	District /City	Poverty Level (%)
1	Nias	17.64
2	Mandailing Natal	10.98
3	Tapanuli Selatan	11.15
4	Tapanuli Tengah	14.58
5	Tapanuli Utara	11.25
6	Toba Samosir	10.08
7	Labuhan Batu	8.95
8	Asahan	11.86
9	Simalungun	10.81
10	Dairi	8.9
11	Karo	9.81
12	Deli Serdang	4.86
13	Langkat	11.36
14	Nias Selatan	18.6
15	Humbang Hasundutan	9.78
16	Pakpak Bharat	10.72
17	Samosir	14.4
18	Serdang Bedagai	9.53
19	Batu Bara	12.24
20	Padang Lawas Utara	10.87
21	Padang Lawas	8.69
22	Labuhan Batu Selatan	11.49
23	Labuhan Batu Utara	10.97
24	Nias Utara	30.92
25	Nias Barat	28.36
26	Kota Sibolga	13.3
27	Kota Tanjung Balai	14.49
28	Kota Pematang Siantar	9.99
29	Kota Tebing Tinggi	11.7
30	Kota Medan	9.3
31	Kota Binjai	6.67
32	Kota Padangsidimpuan	8.32
33	Kota Gunungsitoli	23.43

According to the initial cluster classification of Cluster Hierarchical Cluster Analysis using SPSS, it is known that 17 districts/cities in North Sumatra are districts with low poverty status, while the remaining 16 districts/municipalities are districts with high poverty status.

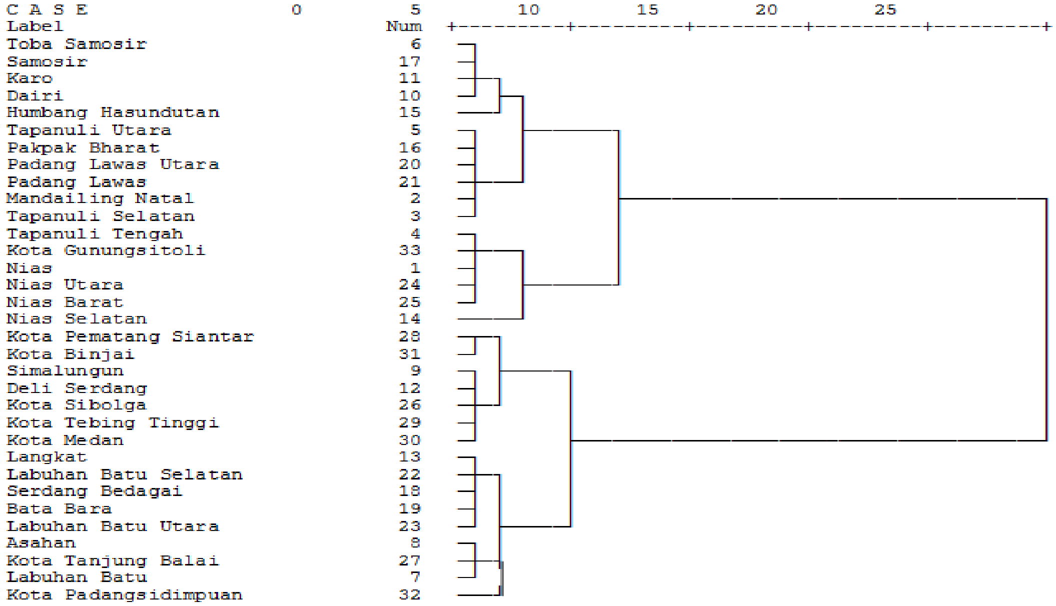

Figure 1. Cluster Hierarchical Cluster Analysis

A. Assumption Testing Discriminant Analysis

1) Multivariate Normal Test

Hypothesis

H_0 : Educational data resolved at primary, junior, secondary, non-working, working in the informal sector, working in the formal sector, users of contraceptives, the percentage of under-fives immunized, households of water users and households own toilet / shared normal multivariate.

H_1 : Educational data resolved at primary, junior, senior secondary, unemployed, informal sector, working in the formal sector, contraceptive users, the percentage of immunized toddlers, households of water users, and households own / shared toxic normal multivariate.

Level of significance ($\propto$) = 0.05

Statistics Count:

p-value = 0.985

Testing Criteria:

H_0 rejected if p-value $\leq \propto$

Decision :

H_0 accepted because p-value = 0.985

B. The Similarity Covariance Matrix Variants Test

1) Hypothesis

H_0 : the covariance matrix of the high poverty status group and the low status group of poverty is the same.

H_1 : the covariance matrix of the high poverty status group and the low status group of poverty is different.

Level of significance ($\propto$) = 0.05

Statistics Count :

$MC^{-1} = 0.08$

Sig. = 0.00

Testing Criteria :

H_0 rejected if Sig. $\leq \propto$

Decision :

H_1 accepted because Sig. = 0.00 $\leq \propto = 0.05$

C. Outlier

The outlier is a datum that deviates very far from the other datum in a sample or collection of datum. Scatter plot indicator Poverty districts / municipal ProvinsI of North Sumatra by using SPSS obtained:

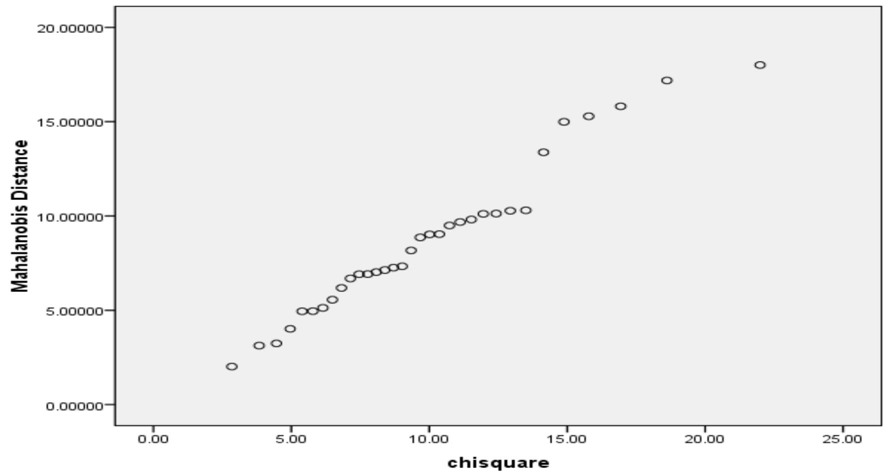

Figure 2. Plot indicators that spread Poverty districts/cities of North Sumatra Province

D. Variable Selection Free

Based on the results of the selection of independent variables using SPSS, the independent variables that influence in determining the status of poverty level of regencies/cities in North Sumatra are variables X_1, X_2, X_3, X_4, X_6, X_7, X_8, X_9, and X_{10}.

E. Classical Linear Discriminant Analysis

The discriminant parameter of discriminant function in classical linear discriminant analysis is done by finding $\bar{x}$ and S which is the average vector and covariance matrix of the data. Based on calculations using standardized data, the average vector and covariance matrices are then converted, so that the classic linear discriminant function is obtained:

$Y = 41{,}13131X_1 - 4810{,}25X_2 - 4771{,}78X_3 - 5{,}94152X_4 - 2080106X_6 + 1683{,}34X_7 - 13{,}0409X_8 + 714{,}8802X_9 + 4093{,}957X_{10}$.

Table 6. Classification results of the classic linear discriminant model for 33 data

Initial Grouping	Grouping By Discriminant Functions		Total
	Group I	Group II	
Group I	14	3	17
Group II	4	12	16
Total	18	15	33

Table 6 explains that many objects classified appropriately for linear discriminant models in 33 regency/city data are as many as 26 objects and many objects are misclassified as many as 7 objects. Based on APER value, the total proportion of classical linear discriminant error is 21.2%, so classical linear discriminant classification is 78.8%.

F. Robust Linear Discriminant Analysis

The estimation of discriminant function parameter in robust linear discriminant analysis is done by replacing $\bar{x}$ and S with $\bar{x}_{MCD}$ and S_{MCD} which is the average vector and covariance matrix with fast-MCD method. Based on calculations using data that has been standardized with Software R, obtained the results of average vekror and covariance matrices are then converted, so that obtained linear discriminant function robust shown as follows:

$Y = -0{,}02655X_1 - 0{,}01483X_2 + 0{,}00864X_3 + 0{,}0025X_4 - 0{,}00924X_6 - 0{,}00683X_7 - 0{,}00448X_8 - 0{,}01477X_9 - 0{,}01078X_{10}$

Table 7. Results Classification of robust linear discriminant model for 33 data

Initial Grouping	Grouping By Discriminant Functions		Total
	Group I	Group II	
Group I	16	1	17
Group II	1	15	16
Total	17	16	33

The above table explains that many of the objects are classified appropriately for robust linear discriminant model data 33 districts/cities are as many as 31 objects and many objects were misclassified as much as two objects. Based on APER value, the total proportion of classical linear discriminant error is 6%, so classical linear discriminant classification is 94%.

IV. CONCLUSION

Classification of poverty data of district/city communities in North Sumatra with classic linear discriminant functions of two groups: $Y = 41{,}13131X_1 - 4810{,}25X_2 - 4771{,}78X_3 - 5{,}94152X_4 - 2080106X_6 + 1683{,}34X_7 - 13{,}0409X_8 + 714{,}8802X_9 + 4093{,}957X_{10}$, resulting in a total proportion of classification mistakes of 21.2%. Classification on poverty data of district/city communities in North Sumatra with linear discriminant function of two groups robust: $Y = -0.02655X_1 - 0{,}01483X_2 + 0{,}00864X_3 + 0{,}0025X_4 - 0{,}00924X_6 - 0{,}00683X_7 - 0{,}00448X_8 - 0.01477X_9 - 0.01078X_{10}$, resulting in a total proportion of classification mistakes of 6%. The linear robust discriminant model classifies the object more precisely than the classic linear discriminant model. This can be seen from the total proportion of classification mistakes of 6%, less than the total proportion of classical linear discriminant classifier error classification of 21.2%. This is due to the large number of outlier in the district/city poverty data in North Sumatra.

REFERENCES

[1] An, J. and Jin, J. 2011. Robust Discriminant Analysis and Its Application to Identify Protein Coding Regions of Rice Genes. Mathematical Biosciences, Vol. 232, 96-100.
[2] Beaver, William H. 1966. Financial Ratios as Prediction Failure. Journal of Accounting Research, Vol. 4, No. 4, pp 71-111.
[3] Bollerslev and Richard T.B. 1994. Cointegration, Fractional Cointegration, and Exchange Rate Dynamics . The Journal of Finance, Vol. 49, No. 2, pp 737-745.
[4] The Central Bureau of Statistics. 2016. National Social Economic Survey 2016. Medan (ID): BPS.
[5] Flury, Bernhard and Riedwyl. 2013. Multivariate Statistics A Practical Approach. London: Chapman and Hall.
[6] Goyal, A and Welch, I. 2003. Predicting the equity premium with dividend ratios. Management Science, Vol. 49, Issue 05, 639 - 654.
[7] Ijumba, Claire. 2013. Multivariate Analysis Of The Brics Financial Markets. South Africa. Pietermaritzburg.
[8] Johnson, Richard A, and Dean. 2007. Applied Multivariate Statistic Analysis. United States of America. Pearson Prentice Hall. Ed ke – 6.
[9] Kerby, April and Alma. 2004. A Multivariate Statistical Analysis of Stock Trends. Miami: University of Miami.
[10] Lange, K and Wu TT. 2008. An MM Algorithm For Multicategory Vertex Discriminant Analysis. J Comput Graph Stat, Vol. 17, N0. 3, 527-544.
[11] Rousseuw, P. J and Driessen, K. V. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics, Vol. 46, No. 3, 293 – 305.
[12] Zohra, Kerroucha Fatima and Bensaid Mohamed. 2015. Using Financial Ratios to Predict Financial Distress of Jordan Industrial Firms: Empirical Study Using Logistic Regression. Academic Journal of Interdisciplinary Studies, Volume 4 No. 2, E-ISSN 2281-612.