Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 4 (2015), pp. 2469-2475

http://www.ripublication.com

Comparison of Vertex Discriminant Analysis (VDA) and Quadratic Discriminant Analysis (QDA)

Helga Kurnia[1], I Made Sumertajaya[2], Farit M. Afendi[3]

Department of Statistics
Bogor Agricultural University
Jalan Pajajaran, Kampus IPB Baranangsiang, Bogor 16151, Indonesia
1.helga.faizin@gmail.com, 2. imsjaya@yahoo.com, 3. fmafendi@gmail.com

Abstract

This study comparestwo different types of the discriminant analysis, namely the vertex(VDA) and quadratic discriminant analysis(QDA) using simulated data. It takes into account the structure of covariance matrix and the ratio of the number of variables to the number of observations. When the number of observations more than number of variables (n > p), the performance of VDA and QDA are relatively similar. It seen from the accuracy of classification. However, when variance between classes is large and distance of mean between classesis near, the VDA's accuracy is lower than the QDA. When the number of observation less than number of variables (n < p), only VDA can be executed. Data from a case study data show the same results as the ones from simulated data.

Keywords: Vertex Discriminant Analysis, Quadratic Discriminant Analysis, Multivariate Analysis

INTRODUCTION

Discriminant analysis is an analysis of multiple variables in the data used to classify each observation into independent groups based on certain variables. Discriminant analysis is still actively progressing. Discriminant analysis which is often used is linear discriminant analysis (LDA) with the Fisher approach. The formation of the LDA discriminant function involves joint covariance matrix components. Joint covariance matrix can be formed if the inter-class covariance matrix structure are same so that it can be combined. When the inter-class covariance matrix is different, the use of LDA become invalid, quadratic discriminant analysis (QDA) can overcome this problem. When the number of variables more than observations (n <p), LDA and

QDA can not be done because the rank of the matrix is smaller than the number of variables, resulting covariance matrix is singular, so it does not haveinverse.

It can be overcome with vertex discriminant analysis (VDA). This research will be carried out comparisons between VDA and QDA.Research on comparisons between the VDA and the LDA has been done by Nurmaleni (2015), that LDA better than VDA on classification ability if distance of mean between classes is near. Perform of LDA and VDA are relatively same in the others conditions. The purpose of this study is to compare VDA with QDA on the data which is n <p and n> p, with n is the number of observations and p is the number of variables.

DATA AND METHODS

Data Simulation

This research use simulation data that generated with n> p as many as 100 repetitions using Cholesky method. For a positive semidefinit symmetric matrix (*A*), can be obtained upper triangular matrix (U) so that $A = U'U$. The equation called Cholesky decomposition (Mattjik and Sumertajaya 2011). Simulation process consists of two stages, namely:

Phase I: Defining the simulation scenario

Simulation scenarios are presented in Table 1. Small differences of variance are determined by insignificant variance, medium differences of variance are determined by variance which began to manifest, while large different of variancedetermined by the difference variance among classeswhich is obvious.

Table 1 Scenario of simulation

Scenario	Variance between classes	Distance of mean between classes
1	Equal	near, medium, long
2	Small	near, medium, long
3	Moderate	near, medium, long
4	Large	near, medium, long

Phase II: Generating Data

Steps on generating data is as follows:

a. Determine the number of groups to be formed, ie 3 groups.
b. Determining the sample size for each class, which is 20.
c. Determine the number of independent variables (X), which is 3 variables (X_1,X_2,X_3).
d. Determine the correlation matrix $\rho = \begin{bmatrix} 1 & 0.1 & 0.4 \\ 0.1 & 1 & 0.8 \\ 0.4 & 0.8 & 1 \end{bmatrix}$.
e. Determining the variance of each class (S_j) in accordance with the simulated scenario specified in phase I.

f. Determining the variance-covariance matrix $\sum$ with formula $\Sigma j = S_j \rho S_j$ for each class.
g. Determining the mean vector for each class in accordance with pre-determined simulated scenarios (μ_1,μ_2,μ_3) is $\mu_A = (3, 6, 9)$, $\mu_B = \mu_A + S_B$, dan $\mu_B = \mu_B + 2S_C$.
h. Generating standard normal random variable Z_j with Z_j~$N_j(0,1)$ by 20 for each class.
i. Generating double normal random variable G_j as many as 20 for each class j, with $G_j \sim Np(\mu_j,\Sigma_j)$ using Cholesky method.
j. Combining all the data classes into one simulated data.

The above steps do as much as 100 repetitions for each scenario using R program.

Method of Analysis

The analysis method used in this study includes the several stages: data exploration, halve data into training and test data, fit the discriminant function in the training data and test data, classify each observation unit into a number of classes.

In the datasimulation, data exploration is done to ensure the generation of data is in conformity with the simulated scenarios, namely: checkthe average of the distance (mean) and the variance between classes, Box's test.

Halve the data into training data and testing data with a ratio of 4: 1. Training data used to establish the discriminant function, and the function that is formed can predict class if the data in every variable of class is entered. So the comparison between prediction class and actual class generates accurat classification. While testing data used to test the accuracy of the classification seemed to determine the class of the new data. Selection of training data and data testing is done randomly.

Fit the discriminant function in the training data, i.e. the method of Vertex Discriminant Analysis (VDA) and Quadratic Discriminant Analysis (QDA).Vertex calculation process on Discriminant Analysis (VDA) is quite complicated when done manually because it involves a process of iteration, and so we need the help of a software program, a program of R with VDA package. QDA calculation process is also done with the help of the program R with the package MASS. VDA algorithm can be written as follows (Wu and Lange, 2008):

i. Determine the initial value of iteration $t = 0$, and initialization $A^{(0)} = 0$ and $b^{(0)} = 0$;
ii. Define $y_i = v_j$ if *i-th* subject to be category of the *j-th*, where
$$v_j = \begin{cases} (k-1)^{-\frac{1}{2}}1, \\ c1 + \text{de}_{j-1}, \end{cases}$$
for $j = 1$
for $2 \leq j \leq k$;
iii. Make major fromobjective function with the *i-th* residual $r_i^{(t)} = y_i - A^{(t)}x_i - b^{(t)}$;

iv. Minimize replacement function and determine $A^{(t+1)}$ dan $b^{(t+1)}$ by the way of completing k set of linear equations.

v. If $\left\|A^{(t+1)} - A^{(t)}\right\| < \gamma$ and $\left|R\left(A^{(t+1)}, b^{(t+1)}\right) - R\left(A^{(t)}, b^{(t)}\right)\right| < \gamma$ both to $\gamma = 10^{-4}$,then stop. Otherwise, repeat step *iii* up to step *v*.

Formation algorithms for discriminant function in QDA are as follows (Johnson and Wichern 2007; Venables and Ripley 2002):

i. Forming a quadratic discriminant function k which is defined the group
$d_j^Q(x) = \ln p_j - \frac{1}{2}(x-\mu_j)'\sum_j^{-1}(x-\mu_j) - \frac{1}{2}\ln\left(\left|\sum_j\right|\right), j = 1,2,...,k$
With μ_j = the average population of the *j-th* grade, *j* = 1,2,...,*k*
p_j = prior probability, if its value is not known, then $p_j = \frac{1}{k}$.
$\sum_j$ = variance-covariance matrix of the *j-th* grade

ii. Due to μ_j and $\sum_j$ is not known, so it can use the estimate of example. Thus, the estimation of a quadratic discriminant scores can be $Q_j(x) = \hat{d}_j^Q(x) = \ln p_j - \frac{1}{2}(x-\bar{x}_j)'S_j^{-1}(x-\bar{x}_j) - \frac{1}{2}\ln\left(\left|S_j\right|\right)$.

Classify each observation unit of training data and test data into a number of classes. In VDA, it is done by selecting the closest distance between the probe with the indicator classes that may (the *j-th* vertex point), with the formula $\hat{y}=\text{argmin}_{j=1,\dots,k}\left\|\boldsymbol{v}_j\text{-}\widehat{\mathbf{A}}\mathrm{x}_i\text{-}\widehat{\mathbf{b}}\right\|$, In QDA,classification rule is to allocate x into class *j* if $Q_i(\boldsymbol{x}) = \text{maks}\left(Q_1(\boldsymbol{x}), Q_2(\boldsymbol{x}), \dots, Q_k(\boldsymbol{x})\right)$.

Calculating the percentage of classification accuracy both the training data and test data. Classification accuracy is the percentage of accuracy between predicted and actual class at all observation units. Classification accuracy is an indicator to see the ability of classification in discriminant analysis method.

Compare the VDA with QDA. Comparisons are based on the ability of better classification. Classification capabilities can be seen from the percentage of accuracy of the classification. In the simulation data, comparison is done based on the variance between classes and / or distance of the midpoint between classes.

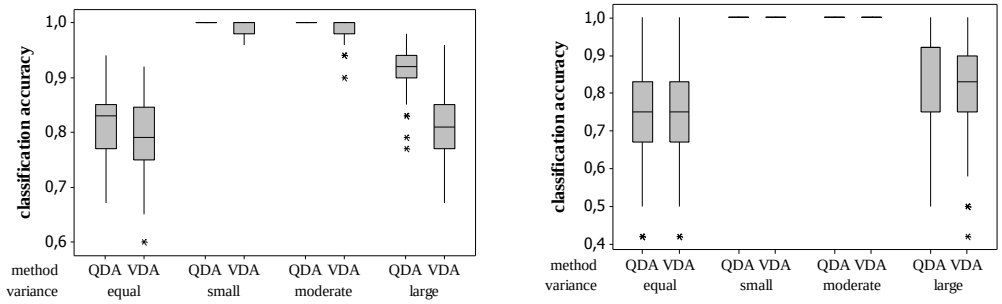
RESULTS AND DISCUSSION

Simulation Study

In the simulation study there were four groups of scenarios. Each group was analyzed using QDA and VDA method for each training data and test data with 100 repetitions, so that each generate 100 classification accuracy. Comparison of classification capabilities between the VDA and QDA is based on the variance between classes and the middle value between classes.

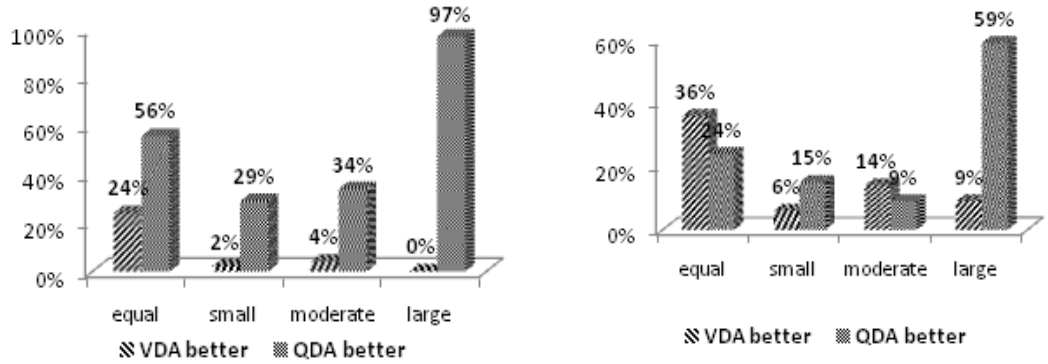
Comparison among 100 classification accuracy between the VDA and QDA based on variance between classes are presented by boxploton the training data and test data in Figure 1. When compared among four scenario variances, VDA and QDA classification capability is very good at small and medium variance, almost no

misclassification. While at the same variance and great variance, accuracy of the classification is more diverse and not as good on small and medium variance. VDA classification capability and QDA is almost the same in general. On the other hand, VDA has less accuracy than QDA at the time of great variance between classes.



a)training data b)test data

Figure1Boxplot classification accuracy of simulation data



a)training data b)test data

Figure 2Comparison of percentage between QDA and VDA classification of 100 repetitions of simulation data

There are three possible conditions at each data simulation replicates, namely: classification accuracy of VDA is greater so that "VDA better", the classification accuracy of QDA is greater so that "QDAbetter", and the classification accuracy of the VDA and QDA are same. Based on these three conditions, the number of "VDA better" and "QDA better" are shown in percentage at each variance of scenario, and both are compared as presented in Figure 2. Figure 2a is a comparison of the training data and Figure 2b is a comparison of test data. Figure 2a shows that the percentage of "QDA better" is greater than "VDA better" for all scenarios variance. However, significant differences occurred only in great variance. On the test data(Figure 2b), the percentage is about the same, except for the great variance, the percentage of "QDA

better" significantly larger.

Comparison results of the average accuracy of classification based on the midpoint between classes and between social variance are presented in Table 2 for training data and Table 3 for test data. Classification capability between the VDA and QDA is almost the same. Except at the midpoint between classes which is close and great variance, QDA superior to VDA. Both tablesshow that the farther the midpoint between classes, classification accuracy will be higher, even up to 100%, which is means that there is no misclassification at all. It is caused when the farther the midpoint between classes, the less likely the occurrence of overlapping boundaries between classes.

Table 2 Comparison of the average accuracy of classification of training data between the mean inter-class populations.

The mean population between classes	equal		small		medium		large	
	QDA	VDA	QDA	VDA	QDA	VDA	QDA	VDA
close	79,8%	76,0%	99,9%	99,0%	99,5%	98,3%	99,0%	84,5%
moderate	92,2%	91,7%	100%	100%	100%	99,9%	88,1%	85,7%
far	99,6%	99,8%	100%	100%	100%	100%	99,9%	99,8%

Table 3 Comparison of the average accuracy of test data classification between the mean inter-class populations

The mean population between classes	equal		small		medium		large	
	QDA	VDA	QDA	VDA	QDA	VDA	QDA	VDA
close	79,8%	76,0%	98,6%	97,4%	97,4%	97,8%	96,5%	82,6%
moderate	92,2%	91,7%	100%	100%	99,6%	99,9%	81,6%	82,2%
far	99,6%	99,8%	100%	100%	100%	100%	100%	99,9%

CONCLUSION

The conclusions obtained from this study are: the data by the number of observations is greater than the number of variables (n> p), classification capabilities of Vertex Discriminant Analysis (VDA) and Quadratic Discriminant Analysis (QDA) is almost the same in general. But VDA has less accuracy than QDA classification at the time of great variance between social classes and distance near the midpoint between classes. On the data with the number of observations is smaller than the number of variables (n <p), only VDA could do the analysis. While QDA can not be done because the rank of the matrix is smaller than the number of variables (p), resulting variance-covariance matrix is singular and has no inverse.

REFERENCES

1. Johnson RA, Wichern DW. 2007. *Applied Multivariate Statistical Analysis*. New Jersey (US): Pearson Prentice Hall. Ed ke-6.
2. Mattjik AA, Sumertajaya IM. 2011. SidikPeubah Ganda. Bogor (ID): IPB Press.
3. Nurmaleni. 2015. Perbandingan metode multikategori *Vertex Discriminant Analysis* dan Analisis Diskriminan Fisher [tesis]. Bogor (ID): Institut Pertanian Bogor.
4. Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. New York (US): Springer.
5. Wu TT, Lange K. 2008. *An MM Algoritm For Multicategory Vertex Discriminant Analysis*. J Comput Graph Stat. 17:527-544.
6. Wu TT, Wu Y. 2012. *Nonlinear Vertex Discriminant Analysis with Reproducing Kernels*. *Statistical Analysis and Data Mining*. doi: 10.1002/sam.11137.