A Model Classification Technique for Linear Discriminant Analysis for Two Groups

Friday Zinzendoff Okwonu [1,2] **and Abdul Rahman Othman** [1]

[1] Univsersiti Sains Malaysia, 11800, Pulau Pinang, Malaysia

[2] Department of Mathematics and Computer Science, Delta State University, P.M.B.1, Abraka, Nigeria

Abstract

Linear discriminant analysis introduced by Fisher is a known dimension reduction and classification approach that has received much attention in the statistical literature. Most researchers have focused attention on its deficiencies. As such different versions of classification procedures have been introduced for various applications. In this paper, we attempt not to robustify the Fisher linear discriminant analysis but to propose a comparable model for dimension reduction and classification. The proposed model is investigated and compared with Nearest mean classifier and Fisher classification rule using unscaled normal and scaled normal generated data. Numerical simulations reveal that the proposed model performed exactly as Fisher's approach and outperformed nearest mean classifier.

Keywords: *Fisher Linear Discriminant Analysis,FZOARO NMC, Classification, Hit-Ratio*

1. Introduction

The sample mean and sample covariance matrix are the corner stone of classical multivariate analysis including linear discriminant analysis[1]. Most statistical experiments are multivariate in nature and large scale multivariate datasets are tractable due to recent technological advancement in computer technology. Classical multivariate analysis relies on the assumption of normality or near-normality, which is often difficult to justify in practice[2]. Linear discriminant analysis (LDA) relies on the sample mean and covariance matrices computed from different groups from the training sample[3]. Linear discriminant analysis is performed using Fisher's technique [3-5]. Fisher's linear discriminant analysis (FLDA) is a linear combination of observed or measured variables that best describe the separations between known groups of observations. Its basic objective is to classify or predict problems where the dependent variables appear in a qualitative form [6-9]. Fisher linear discriminant analysis is a conventional multivariate technique for dimension reduction and classification [10-12]. FLDA often enjoy basic advantages such as robustness to non-normality and even to mildly variation in covariance matrices[10, 13]. The basic disadvantages of FLDA are that it applies same covariance matrix for all groups. It also require large training sample size for good generalization[14, 15]. The Nearest Mean Classifier (NMC) is a simple classifier technique which can provide state of the art performance in relatively high dimensional data[16-18]. [16] observed that NMC may be seen as a basis for LDA and its penalized and flexible variations. In this paper, we proposed a comparable classification model to the classical FLDA and compare its classification competency with FLDA and NMC. The proposed approach is not to undermine FLDA but as an alternative classification and dimension reduction technique.

This paper is organized as follows. In section two, we describe Fisher linear discriminant analysis. FZOARO classification model is proposed in section three. Nearest mean classifier is briefly described in section four. Simulations and conclusions are presented in sections five and six.

2. Fisher Linear Discriminant Analysis

Fisher suggested transforming multivariate observations to univariate observations such that the univariate observations derived from each population is maximally separated. The separation of these univariate observations can be examined by their mean difference[13]. Fisher classification rule maximizes the variation between samples variability to within samples variability. Fisher linear discriminant analysis for two groups is described as follows.

Consider classifying an observation vector $\mathbf{x}$ into one of two populations say $\pi_i : N_p(\mu_i, \Sigma), (i = 1, 2)$, the population mean vectors and covariance matrix are denoted as μ, Σ respectively. Since the population mean vectors and covariance matrix are unknown, the sample estimate is used in this paper. We define the sample mean vectors, sample covariance matrix and pooled sample covariance matrix as follow;

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij},$$

$$S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\right)',$$

$$S_{pooled} = \frac{\sum_{i=1}^{g=2}(n_i - 1)S_i}{\sum_{i=1}^{g=2} n_i - g}.$$

Applying the parameters define above, Fisher linear discriminant analysis[19] can be started as;

$$Z = C'X = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' S_{pooled}^{-1} X, \qquad (1)$$

$$\bar{Z} = (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)C/2. \qquad (2)$$

Eq. (1) and (2) are the discriminant score and discriminant mean. The classification rule based on eq. (1) and (2) can be described as follows.

Allocate $\mathbf{x}_1$ to population one π_1 if

$$Z \geq \bar{Z}, \qquad (3)$$

otherwise allocate $\mathbf{x}_1$ to population two π_2 if

$$Z < \bar{Z}. \qquad (4)$$

Fisher maintained that his technique most adhere strictly to equal variance covariance matrix of the two normal populations. FLDA has been extended to more than two groups[19, 20].

3. FZOARO Classification Model

In this section, we propose a comparable classification and dimension reduction technique called FZOARO classification model. The proposed technique is developed based on the following assumptions; the sample size most be greater than the sample dimension, and secondly, the equal variance covariance matrix most come from a multivariate normal distribution. Since the population parameters are unknown, we use the sample equivalent. Let the sample mean vectors, sample covariance matrix and pooled sample covariance matrix be define as follow;

$$\bar{\mathbf{x}}_{emui} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{obsij},$$

$$S_{oroi} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(\mathbf{x}_{obsij} - \bar{\mathbf{x}}_{emui}\right)\left(\mathbf{x}_{obsij} - \bar{\mathbf{x}}_{emui}\right)',$$

$$S_{pooled} = \frac{\sum_{i=1}^{g=2}(n_i - 1)S_{oroi}}{\sum_{i=1}^{g=2} n_i - g}.$$

Applying the above sample parameters, we state the FZOARO classification model mathematically as follow;

$$\Omega_{emonu} = \eta_\omega \mathbf{X} = (\mathbf{w}'\delta_{orogu})\mathbf{X}, \qquad (5)$$

where

$$\eta_\omega = \mathbf{w}'\delta_{orogu},$$

$$\mathbf{w} = \left(\bar{\mathbf{x}}_{emu1} - \bar{\mathbf{x}}_{emu2}\right)' S_{pooled}^{-1},$$

$$\delta_{orogu} = \left(\sum S_{pooled}^{-1} \Big/ \sqrt{\chi_p^2}\alpha = h\right),$$

$$h = 3n/4n,$$

$$\bar{\Omega}_{emonu/orogu} = (\bar{\mathbf{x}}_{emu1} + \bar{\mathbf{x}}_{emu2})\eta_\omega / 2. \qquad (6)$$

The parameter Ω_{emonu} is the discriminant score and $\bar{\Omega}_{emonu/orogu}$ is the discriminant mean. The classification procedure is performed by comparing the discriminant score with the discriminan mean. That is,.

Classify $\mathbf{x}_{obs1}$ to population one ϖ_1 if

Ω_{emonu} is less than or equal to $\bar{\Omega}_{emonu/orogu}$

or classify $\mathbf{x}_{obs1}$ to population two ϖ_2 if

Ω_{emonu} is less than $\bar{\Omega}_{emonu/orogu}$.

This approach is relatively computationally simple. Note that eq. (6) contains a constant δ_{orogu} which requires Chi-square value. This constant stabilizes the discriminant coefficient against boundary observations, in other words, it help to adjust classification accuracy of observations that are nearer or far from the decision boundary. Put differently, this constant is considered to have dual functions one as a stabilizer and two as control factor.

4. Nearest Mean Classifier

The nearest mean classifier estimates the mean of every group and allocates new observation samples to the corresponding group with the nearest mean. The nearest mean classifier (NMC)[21] is started as

$$Z_{NMC} = \mathbf{dX} = (\mathbf{x}_1 - \mathbf{x}_2)'\mathbf{X}, \quad (7)$$
$$\mathbf{d} = (\mathbf{x}_1 - \mathbf{x}_2)',$$

$$\bar{Z}_{NMC} = \frac{(\mathbf{x}_1 + \mathbf{x}_2)\mathbf{d}}{2}. \quad (8)$$

Eq. (7) and (8) are the building block of NMC approach. The classification approach follow the same procedure as explained in section two, eq. (3) and (4). Detail of this approach can be found in [17, 18, 22].

5. Simulation

In this experiment we investigate the comparative classification performance of the three classification procedures discussed in this paper. The objective of this paper is to obtain comparable classification procedure to Fisher's approach and to compare it with nearest mean classifier. The sample mean and covariance matrix are fixed. The sample sizes for both groups are equal $n_1 = n_2 = 100, N = n_1 + n_2$. The numerical simulation is conducted using unscaled and scaled normal generated data. The data are randomly selected and is divided into two; say 70% training and 30% validation. The experiment was run 1000 times and the average number of runs reported. Result from table 1 shows that our model performs exactly as Fisher LDA and both outperform nearest mean classifier (NMC). The numbers in bold face are the observations that were correctly classified while numbers in bracket italics are the percentage of observations that were misclassified. Results from table 1 reveal that the proposed model and FLDA perform similar and both techniques outperform NMC for the unscaled data. Table 2 affirms the result in table 1.

Table 1: Unscaled Normal Data

Methods/Hit-ratio (%)		
FLDA	FZOARO	NMC
99.7(*0.3*)	**99.7**(*0.3)*	**99.5**(*0.5*)

Table 2: Scaled Normal Data

Methods/Hit-ratio (%)		
FLDA	FZOARO	NMC
60.9(*39.1*)	**60.9**(*39.1*)	**60.2**(*39.8*)

6. Conclusions

We have proposed a comparable classification and dimension reduction technique to FLDA. In this paper, we performed numerical simulations using generated data to validate the competency of our approach. The new approach is compared with FLDA and NMC. Simulation results show that FZOARO perform exactly as FLDA and outperformed NMC. The proposed approach is not intended to invalidate FLDA however it can be used as an alternative method for classification and dimension reduction.

References

[1] Y. Zuo, " Robust Location and Scatter Estimators in Multivariate Analysis," *WSPC/Trim Size:9in x6in for Review Volume,*0-31, 2005.

[2] R. Y. Liu; J. M. Parelius; and K. Singh, "Multivariate Analysis by data depth: Discriptive Statistics, Graphics and Inference," *The Annals of Statistics*, vol. 27,783-858, 1999.

[3] C. Croux.; P. Filzmoser ; & K. Joossens, "Classification Efficiency for Robust Linear

Discriminant Analysis," *Statistica Sinica,* vol. 18, pp. 581-599, 2008.

[4] M. Sever; J. Lajovic; and B. Rajer, "Robustness of the Fisher's Discriminant Function to Skew-Curved Normal Distribution," *Metodoloski zvezki,* vol. 2, pp. 231-242, 2005.

[5] D. M. Witten; and R. Tibshirani, "Penalized classification using Fisher's linear discriminant," *J. R. Statist. Soc. B,* vol. 75, pp. 753-772, 2011.

[6] A. C. Rencher, "Methods of Multivariate Analysis," second, Ed.: A John Wiley & Sons, Inc., 2002.

[7] R. A. Johnson; & D. W. Wichern, "Applied Multivariate Statistical Analysis, Prentice -Hall International, Inc., 1998.

[8] E. I. Altman, "Financial Ratios, Discriminant Analysisand the Prediction of Corporate Bankruptcy," *The Journal of Finance,* vol. 23, pp. 589-609, 1968.

[9] M. Hubert ;and K. V. Driessen "Fast and Robust Discriminant Analysis," *Comput. Statist. Data Anal.,* vol. 45, pp. 301-320, 2004.

[10] T. Hatie; A. Buja; and R. Tibshirani, "Penalized Discriminant Analysis," *The Annals of Statistics*, vol. 23, pp. 73-102, 1995.

[11] R. J. Durant; and A. Kaban, "Compressed Fisher linear discriminant analysis:classification of randomly projected data," *KDD'10,Washington DC, USA*, pp. 1-10, 2010.

[12] M. Pohar; M. Blas; and S. Turk, "Comparison of Logistic Regression and Linear Discriminant Analysis:A Simulation Study," *Metodoloski zvezki,* vol. 1, pp. 143-161, 2004.

[13] R. A. Johnson; & D. W. Wichern, "Applied Multivariate Statistical Analysis,: Prentice HaLL International Editions, 2002.

[14] W. Deng; J. Hu; and J. Guo, "Robust Fisher Linear Discriminant Model for Dimensionality Reduction," *IEEE,* pp. 1-4, 2006.

[15] M. Loog; and R. P.W. Duin. "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff Criterion," *IEEE Trans. PAMI,* vol. 26, pp. 732-739, 2004.

[16] M Loog, "Constrained Parameter Estimation for Semi-supervised Learning:The Case of the Nearest Mean Classifier," *ECML PKDD 2010,Part II, LNAI 6322,* pp. 291-304, 2010.

[17] G. Mclachlan, "Discriminant analysis and statistical pattern recognition," John Wiley & Sons, Chichester, 1992.

[18] R. Duda; and P. Hart, "Pattern classification and scence analysis," John Wiley &Sons, Chichester, 1973.

[19] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics,* vol. 7, pp. 179 - 188, 1936.

[20] C. R. Rao, "The ultilization of multiple measurements in problems of biological classification," *Journal of royal statist. soc.,*, vol. 10, pp. 159-193, 1948.

[21] M. Skurichina; and R. P. W. Duin, "Boosting in linear discriminant analysis," *www.tnw.tudelft.nl/live/binaries/...6bb6.../mcs_2000_**boosting**.pdf*

[22] K. Fukunaga, "Introduction to Statistical Pattern Recognition," 2 ed: Academic Press Professional, Inc. San Diego, CA, USA,1990.

FRIDAY ZINZENDOFF OKWONU has received M.Sc degree in Mathematics from the University of Ibadan in 2008, Nigeria. M sc in applied Mathematics from Universiti Sains Malaysia in 2011. He is currently a PhD student at Universiti Sains Malaysia. His current research interest is in robust discriminant analysis. He has published and attended so many conferences within and outside Malaysia.

ABDUL RAHMAN OTHOMAN he is a professor of robust statistics. He is a lecturer at the Universiti Sains Malaysia. He has published so many journals including local and international.