Robust Linear Discriminant Analysis with Automatic Trimmed Mean

Sharipah Soaad Syed Yahaya, Yai-Fung Lim, Hazlina Ali, Zurni Omar
School of Quantitative Sciences, UUM College Arts & Sciences, 06010 Universiti Utara Malaysia Sintok, Kedah, Malaysia.
sharipah@uum.edu.my

***Abstract*—Linear discriminant analysis (LDA) is a multivariate statistical technique used to determine which continuous variables discriminate between two or more naturally occurring groups. This technique creates a linear discriminant function that yields optimal classification rule between two or more groups under the assumptions of normality and homoscedasticity. Nonetheless, the computation of parametric LDA which are based on the sample mean vectors and pooled sample covariance matrix are known to be sensitive to non-normality. To overcome the sensitivity of this method towards non-normality as well as homoscedasticity, this study proposed a new robust LDA method. Through this approach, an automatic trimmed mean vector was used as a substitute for the usual mean vector in the parametric LDA. Meanwhile, for the covariance matrix, this study introduced a robust approach by multiplying the Spearman's rho with the corresponding robust scale estimator used in the trimming process. Simulated and real financial data were used to test the performance of the proposed method in terms of misclassification rate. The results showed that the new method performed better compared to the parametric LDA and the existing robust LDA with *S*-estimator.**

***Index Terms*—Linear Discriminant Analysis; Misclassification Rates; Robust; Trimmed Mean.**

I. INTRODUCTION

Linear discriminant analysis (LDA) is a multivariate classification technique to determine which variable discriminates between two or more classes, and to construct a classification model for predicting the group membership of new observations. In short, LDA aims for reliable group allocations of new observations based on a discriminant rule which is developed from a training data set with known group memberships. LDA are known to perform optimally when the assumptions of normality and homoscedasticity are met [1]. However, optimality is hard to achieve as its computation rely heavily on the sample mean vectors and pooled sample covariance matrix. These two statistics are known to be sensitive to outliers, which consequently may increase misclassification rate [2]. To overcome this sensitivity problem in the parametric LDA, researchers seek for alternatives in robust linear discriminant analysis (RLDA). By substituting the classical estimators with robust estimators such as *M*–estimators, Minimum Covariance Determinant (MCD) [3, 4], Minimum Volume Ellipsoid (MVE) [5], and *S*-estimators [6, 7, 8], robust discriminant model with minimum classification error rate could be developed [1].

In this paper, an approach using automatic trimmed mean is proposed in the construction of new RLDA models. Unlike the usual trimming process, the trimming employed in this work take into consideration the distributional shape of the data. Through this trimming approach, only outliers will be trimmed away leaving just the good data. Simulation and real financial data were used to investigate on the performance of the proposed RLDA. For the real financial data, the investigation emphasizes on classifying "distress" and "non-distress" banks in Malaysia. Due to the nature of the real data problem, our work only focuses on two populations. The proposed RLDA were then compared to the classical LDA and also to the existing robust LDA with *S*-estimators. The performance of the discriminants rules were evaluated by misclassification rate provided by simulation and real life study.

II. DISCRIMINANT RULES

Suppose that we have one group of p-dimensional feature data, $\mathbf{x_1}$, from population π_1 of H_1 distribution with mean $\boldsymbol{\mu_1}$ and covariance matrix $\boldsymbol{\Sigma_1}$, and the other group of data, $\mathbf{x_2}$, from population π_2 of H_2 distribution with mean $\boldsymbol{\mu_2}$ and covariance matrix $\boldsymbol{\Sigma_2}$. A discriminant rule can be constructed to assign one new observation $\mathbf{x_0}$ to π_1 or π_2. One of the familiar models to unravel this problem is via classical LDA which is derived under the assumptions that all the populations have identical covariance, such that $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$. The classical discriminant rule is defined as follows in equation (1) [9].

$$\begin{aligned} &\text{If} \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \, \boldsymbol{\Sigma}^{-1}\left[\mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right] \geq \ln\left(\frac{p_2}{p_1}\right) \\ &\text{then} \quad \mathbf{x}_0 \in \pi_1, \\ &\text{else} \quad \mathbf{x}_0 \in \pi_2. \end{aligned} \tag{1}$$

where p_1 and p_2 are the prior probability that an individual comes from population π_1 and π_2 respectively. Practically, the overall misclassification probability can be minimized based on this classical discriminant rule. Since the classical parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, are usually undefined, hence we need to estimate the parameters from the sample data. However, the performance of the classical discriminant rule will be badly affected if non-normality and/or heteroscedasticity occur [10]. It is clear that the classical discriminant rule will become non-robust due to the sensitivity of classical estimates.

By plugging robust estimators for the location, **μ** and scatter, **Σ**, a robust discriminant rules can be developed. The location estimator in this paper is the automatic trimmed mean proposed by Keselman [11]. Trimming is one of the strategies to deal with outliers. This automatic trimmed mean is derived using data left from empirically determined trimming. It is a highly robust location estimator which possesses highest breakdown point and is defined as equation (2).

$$\hat{\boldsymbol{\mu}}_{jk} = \sum_{i=i_1+1}^{n_{jk}-i_2} \frac{\mathbf{x}_{(i)jk}}{n_{jk}-i_1-i_2} \qquad j = 1,\ldots,p;\ \ k = 1,2 \tag{2}$$

where

i_1, i_2 = number of trimmed obs. for the both end of data

$i_1 = x_{(i)jk} \ni \left(x_{(i)jk} - \hat{M}_{jk}\right) < -2.24\left(MADn_{jk}\right)$

$i_2 = x_{(i)jk} \ni \left(x_{(i)jk} - \hat{M}_{jk}\right) > 2.24\left(MADn_{jk}\right)$

$\hat{M}_{jk}$ = median in dimension j for group k

$\mathbf{x}_{(i)jk}$ = i^{th} ordered obs. dimension j for group k

n_{jk} = total number of obs. in dimension j for group k

$MADn_{jk} = 1.4826 \text{ Median}\left\{\left|x_{(1)jk} - \hat{M}_{jk}\right|,\ldots,\left|x_{(n)jk} - \hat{M}_{jk}\right|\right\}$

Meanwhile, the covariance (scatter) matrix for the RLDA is estimated using the product of spearman correlation coefficient (ρ) and rescaled median absolute deviation (*MADn*) as in equation (3).

$$\hat{\boldsymbol{\Sigma}}_k = \begin{bmatrix} MADn_{1k}^2 & \cdots & \rho_{1pk} MADn_{1pk}^2 \\ \vdots & \ddots & \vdots \\ \rho_{p1k} MADn_{p1k}^2 & \cdots & MADn_{pk}^2 \end{bmatrix} \tag{3}$$

The two robust statistics (location and scatter) which replaced the mean and covariance matrix, when paired together in LDA formed a new robust discriminant rule denoted as RLDA_T.

III. Simulation Study

A simulation study was conducted to evaluate on the performance of the proposed RLDA technique, denoted as RLDA_T. These techniques were then compared against the classical LDA and RLDA with *S*-estimators (RLDA_S). To check on the strength and weakness of the existing and the new techniques, a few variables were manipulated to create conditions commonly encountered in real life. These variables were percentage of contamination (ε = 0, 0.1, 0.2), sample sizes (n = 20, 50, 100), shift in location (μ = 0, 5) and shift in shape (κ = 0, 25).

The procedure started by generating a training data set based on the various conditions to develop a discriminant rule for each condition. Next, generate another data set of size 2000 for both groups from uncontaminated populations to validate the corresponding discriminant rules. This experiment is replicated 2000 times for each condition. The performance of the investigated techniques which was based on misclassification rates is presented in Table 1. The digits in brackets are the computational time (in seconds) to compute each condition.

Table 1
Mean and Computational Time of the Misclassification Rate for Various LDA Models

ε	μ	κ	$n_1 = 20$ $n_2 = 20$			$n_1 = 50$ $n_2 = 50$			$n_1 = 100$ $n_2 = 100$		
			LDA	RLDA_S	RLDA_T	LDA	RLDA_S	RLDA_T	LDA	RLDA_S	RLDA_T
0	0	0	0.2115 (3)	0.2126 (1221)	0.2187 (9)	0.2001 (3)	0.2005 (1177)	0.2033 (9)	0.1968 (3)	0.1970 (1393)	0.1985 (9)
0.1	5	0	0.5001 (3)	0.2168 (1231)	0.2492 (9)	0.4993 (3)	0.2013 (1264)	0.2188 (9)	0.5017 (3)	0.1971 (1346)	0.2072 (10)
0.2	5	0	0.6185 (3)	0.5808 (1250)	0.3184 (8)	0.6650 (3)	0.6138 (1278)	0.2723 (9)	0.7020 (3)	0.6478 (1218)	0.2427 (9)
0.1	0	25	0.3719 (3)	0.2131 (1131)	0.2191 (9)	0.3505 (3)	0.2010 (1274)	0.2039 (9)	0.3051 (3)	0.1971 (1323)	0.1985 (9)
0.2	0	25	0.4442 (3)	0.2174 (1134)	0.2209 (9)	0.4074 (3)	0.2022 (1273)	0.2046 (9)	0.3651 (3)	0.1977 (1305)	0.1989 (9)
0.1	5	25	0.4291 (3)	0.2131 (1195)	0.2195 (9)	0.4686 (3)	0.2009 (1242)	0.2040 (9)	0.4829 (3)	0.1971 (1302)	0.1987 (10)
0.2	5	25	0.5295 (3)	0.2176 (1153)	0.2232 (8)	0.5814 (3)	0.2021 (1101)	0.2052 (9)	0.6369 (3)	0.1977 (1338)	0.1993 (9)

The results reveal that all the techniques perform equally well when there is no contamination (third row). Theoretically, under ideal condition, that is when all the assumptions are fulfilled, classical LDA should perform optimally and the results concur with the theory. Nevertheless, the two robust techniques do not perform much worse than the classical LDA. In contrast, when there is contamination (ε), the results show that the misclassification rate for the classical LDA inflates above the other two techniques. At 10% contamination, regardless of the shift in location and shape, RLDA_S performs better than RLDA_T, but the disparities between the two techniques are quite small. When contamination increases to 20%, combined with shift in location, the misclassification rates for RLDA_T are very much smaller than RLDA_S, not to mention the LDA. When the 20% contamination combined with shift in shape, but without shift in location, the rates for RLDA_S and RLDA_T can be interpreted as almost the same for larger sample sizes. For

small sample size, $RLDA_S$ performs slightly better than $RLDA_T$. As the sample size increases, $RLDA_S$ outperforms $RLDA_T$ even though the misclassification rates for $RLDA_T$ decreases. Across the table, we can observe that the misclassification rates for $RLDA_T$ are consistently small, ranging from 19.85% to 31.84% as compared to $RLDA_S$ with the range of 19.70% to 64.78%. Meanwhile, the range for the classical LDA is 19.68% to 70.20%. In addition, the misclassification rates for $RLDA_T$ are consistently improving as the number of sample sizes increases but the pattern does not exist in the other two techniques. Another added value for $RLDA_T$ is the computing time. As shown in the brackets under each condition, the computational time for $RLDA_T$ is very much smaller than $RLDA_S$. Even though the computational time for LDA is consistently smaller than $RLDA_T$, the high misclassification rates when contamination occurs indicate that LDA is not a robust technique and we have to employ it with care.

IV. Real Data Application

Besides simulation study, all the models were also being put to test on real data, specifically, to classify financially distressed and non-distressed banking institutions in Malaysia. The bank data were extracted from selected balance sheet in annual report of 27 commercial banks from year 1988 to 1999. Two independent variables were used to capture variation in financial crisis. The variables were ratio of total shareholder's fund to total assets (CA), and ratio of total shareholder's fund to total equity (EQ). Table 2 shows the results of Lilliefor normality test for both variables in each group.

Table 2
Results of the Lilliefor Normality Test

Group	*p*-value	
	CA	EQ
Distress	0.0066	0.0214
Non-distress	0.1321	0.0011

Normality checking on the financial data showed a violation of normality assumption. The performance of each model was based on its corresponding apparent error rates (AER) and estimate of misclassification rates using cross-validation (CV). The results of the real data analysis are presented in Table 3.

Table 3
Misclassification Rate for the Classical LDA and RLDA

LDA Estimators	AER	CV
LDA	0.1111	0.1111
$RLDA_S$	0.0741	0.1111
$RLDA_T$	0.0370	0.0741

The real data results reveal that all RLDA are able to detect outliers and produces smaller error rates than the classical LDA. However, among the RLDA, the proposed technique ($RLDA_T$) produces smallest error rate as compared to the existing $RLDA_S$. The proposed model is found to be the best as its produces the smallest error rates via AER as well as CV. The simulation and real life problem results proven that the proposed $RLDA_T$ technique provides a comparable performance or better among the investigated LDA.

V. Conclusion

This paper presents an automatic trimmed mean paired with robust covariance to alleviate the classification problem. The outliers were eliminated via trimming process which took into consideration distributional shape of the data before developing the robust discriminant rule. Their function (robust estimators) as substitutes for the classical estimators in linear discriminant analysis (LDA) technique very much improves the misclassification rates. Even when compared to the existing robust LDA using *S*-estimator, the simulation and real data analysis prove that the proposed technique is comparable or sometimes better. The proposed technique produces low error rates as well as computational time. Generally, we can conclude that the robust linear discriminant analysis proposed in this paper should be considered in solving classification problems especially when non-normality (outliers' existence) is suspected.

Acknowledgment

The authors would like to acknowledge the work that led to this paper, which was fully funded by the Fundamental Research Grant Scheme (S/O 12801) of Ministry of Higher Education.

References

[1] C. Croux, P. Filzmoser and K. Joossen, "Classification Efficiencies for Robust Linear Discriminant Analysis," *Statistica Sinica.*, vol. 18, no. 2, pp. 581-599, Apr. 2008.

[2] T.T. Sajobi, L.M. Lix, B.M. Dansu, W. Laverty and L. Li, "Robust Descriptive Discriminant Analysis for Repeated Measures Data," *Computational Statistic and Data Analysis*, vol. 56, no. 9, pp. 2782-2794, Mar. 2012.

[3] M. Hubert and K. Driessen, "Fast and Robust Discriminant Analysis," *Comput. Statist. Data Anal.*, vol. 45, pp. 301–320, Mar. 2004.

[4] M.J. Alrawashdeh, S.R. Muhammad Sabri and M.T. Ismail, "Robust Linear Discriminant Analysis with Financial Ratios in Special Interval," *Applied Mathematical Sciences*, vol. 6, no. 121, pp. 6021-6034, Jun. 2012.

[5] C.Y. Chorl and P.J. Rousseeuw, "Integrating a High-Breakdown option into Discriminant Analysis in Exploration Geochemistry," *Journal of Geochemical Exploration*, vol. 43, no. 3, pp. 191-203, Jun. 1992.

[6] X. He and W.K. Fung, "High Breakdown Estimation for Multiple Populations with Applications to Discriminant Analysis," *J. Multivariate Anal.*, vol. 72, no. 2, pp. 151–162, Feb. 2000.

[7] C. Croux and C. Dehon, "Robust Linear Discriminant Analysis using S-estimators," *Canad. J. Statist.*, vol. 29, no.3, pp. 473–493, Sept. 2001.

[8] Y.F. Lim, S.S. Syed Yahaya, F. Idris, H. Ali and Z. Omar Z, "Robust Linear Discriminant Models to solve Financial Crisis in Banking Sector," in *Proceedings of the 3rd International Conference on Quantitative Sciences and Its Applications*, AIP Conference Proceedings 1635, Kedah, 2014, pp: 794–798.

[9] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall International Edition, New Jersey, 2002, pp: 581-612.

[10] R.M. GlèlèKakaï, D. Pelz and R. Palm, "On the Efficiency of the Linear Classification Rule in Multi-group Discriminant Analysis," *African Journal of Mathematics and Computer Science Research*, vol. 3, no. 1, pp. 019-025, Jan. 2010.

[11] R.R. Wilcox and H.J. Keselman. "Repeated Measures ANOVA Based on a Modified One-Step M-Estimator," *Journal of British Mathematical and Statistical Psychology*, vol. 56, no. 1, pp. 15 – 26, May 2003.