Journal of Mathematics and Statistics

Original Research Paper

Robust Linear Discriminant Analysis

Sharipah Soaad Syed Yahaya, Yai-Fung Lim, Hazlina Ali and Zurni Omar

School of Quantitative Sciences,
UUM College of Arts and Science 06010 Universiti Utara Malaysia, Sintok, Kedah Malaysia

Article history
Received: 22-06-2016
Revised: 28-07-2016
Accepted: 30-12-2016

Corresponding Author:
Yai-Fung Lim
School of Quantitative Sciences, UUM College of Arts and Science 06010 Universiti Utara Malaysia, Sintok, Kedah Malaysia
Email: lim_yai@ahsgs.uum.edu.my

Abstract: Linear Discriminant Analysis (LDA) is the most commonly employed method for classification. This method which creates a linear discriminant function yields optimal classification rule between two or more groups under the assumptions of normality and homoscedasticity (equal covariance matrices). However, the calculation of parametric LDA highly relies on the sample mean vectors and pooled sample covariance matrix which are sensitive to non-normality. To overcome the sensitivity of this method towards non-normality as well as homoscedasticity, this study proposes two new robust LDA models. In these models, an automatic trimmed mean and its corresponding winsorized mean are employed to replace the mean vector in the parametric LDA. Meanwhile, for the covariance matrix, this study introduces two robust approaches namely the winsorization and the multiplication of Spearman's rho with the corresponding robust scale estimator used in the trimming process. Simulated and real financial data are used to test the performance of the proposed methods in terms of misclassification rate. The numerical result shows that the new method performs better if compared to the parametric LDA and the robust LDA with *S*-estimator. Thus, these new models can be recommended as alternatives to the parametric LDA when non-normality and heteroscedasticity (unequal covariance matrices) exist.

Keywords: Discriminant Analysis, Classification, Normality, Homoscedasticity, Robust, Trimmed Mean, Winsorized

Introduction

Linear Discriminant Analysis (LDA) is one of the most widely used statistical approaches for analyzing attribute variables in supervised classification (Elizabeth and Andres, 2012). The purpose of LDA is 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 (equal covariance matrices) are met (Croux *et al*., 2008). However, the high dependencies of its calculation on sample mean vectors and pooled sample covariance matrix may increase misclassification rate in the existence of outliers (Sajobi *et al*., 2012). It is a known fact that mean, which possesses zero breakdown point, is very sensitive to outliers. To overcome this sensitivity problem in 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) (Hubert and Driessen, 2004; Alrawashdeh *et al*., 2012), Minimum Volume Ellipsoid (MVE) (Chorl and Rousseeuw, 1992) and *S*-estimators (He and Fung, 2000; Croux and Dehon, 2001; Lim *et al*., 2014), we can develop robust discriminant model with minimum classification error rate (Croux *et al*., 2008).

In this study, two approaches, namely trimming and winsorizing are proposed in the construction of new RLDA models to create discriminant rule that are robust to deviation. The coordinate-wise based estimators have been applied in this research with the purpose of producing at least one successful RLDA models to solve classification problems. Unlike the usual trimming and winsorizing process, the trimming employed in our work take into consideration the shape of data distribution.

Through this trimming approach, only outliers will be trimmed away leaving just the good data. A simulation study and a real life financial data are used to investigate the performance of the proposed RLDA. We are interested to classify "distress" and "non-distress" banks in Malaysia for the real life financial problem. Therefore, our work is scoped to two populations only due to the nature of the real life problem. The proposed RLDA are then compared to the classical LDA and also to the well-known robust LDA with *S*-estimators. The performance of the discriminants rules are evaluated by misclassification rate provided by simulation and real life study.

The rest of this paper is structured as follows. Section 2 describes about discriminant rules for classical LDA and proposed RLDA. The results and discussions based on the simulation study and real life problem application will be delivered in the section 3. Lastly, the concluding remark will be provided in section 4.

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, 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 Equation 1 (Johnson and Wichern, 2002):

$$\begin{array}{ll} \textit{if} & (\mu_1\text{-}\mu_2)'\Sigma^{-1}\left[x_0-\frac{1}{2}(\mu_1+\mu_2)\right]\geq \ln\left(\frac{p_2}{p_1}\right) \\ \textit{then} & x_0\in\pi_1 \\ \textit{else} & x_0\in\pi_2 \end{array} \qquad (1)$$

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 (unequal covariance matrices) occur (GlèlèKakaï *et al.*, 2010). 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, $\boldsymbol{\mu}$ and scatter $\boldsymbol{\Sigma}$, a robust discriminant rules can be developed.

In this study, we introduced two robust estimators namely the automatic trimmed mean, which is also known as modified one-step *M*-estimator (MOM) and its winsorized version, referred to as winsorized modified one-step *M*-estimator (WMOM) to construct RLDA model. Trimming and winsorizing are among the strategies to deal with extreme values. MOM estimate of location is modified from the one-step *M*-estimator which was introduced by Wilcox and Keselman (2003). Based on the concept of trimmed means, the MOM estimator is derived using data left from empirically determined trimming. Briefly, MOM estimator is a highly robust location estimator which possesses highest breakdown point and is defined as Equation 2:

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

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*
njk = total number of obs. in dimension *j* for group *k*
$MADn_{jk}$= 1.4826 Median $\left\{\left|x_{(1)jk}-\hat{M}_{jk}\right|,\dots,\left|x_{(n)jk}-\hat{M}_{jk}\right|\right\}$

Another strategy in dealing with extreme values is the winsorization approach. Winsorization follows the process of trimming, but instead of discarding the trimmed values, they are replaced by the remaining highest and lowest values. Winsorized MOM (WMOM) follows the same trimming process as MOM before replacing the trimmed values with the remaining lowest and highest end of the data (Haddad *et al.*, 2013). Unlike MOM, WMOM will retain the original sample size and this approach can reduce the problem of losing information due to trimming. WMOM estimate of location and scatter can be defined as Equation 3:

$$\hat{\mu}_k=\sum_{i=1}^{n_{jk}}\frac{W_{ij}}{n_{jk}} \qquad j=1,\dots,p;\quad k=1,2 \qquad (3)$$

where W_{ij} is the winsorization of a random sample.

Meanwhile, the covariance matrix will be estimated using two approaches; the winsorized covariance and the product of spearman correlation coefficient and rescaled Median Absolute Deviation (*MADn*). These covariance matrices will be paired with the corresponding WMOM and MOM location estimates to form robust discriminant rule denoted as RLDA_{WM} and RLDA_{M}, respectively.

Results and Discussions

In this section, simulation study and real data application are implemented to evaluate on the performance of the two proposed RLDA models. These models will then be compared against the classical LDA model and existing RLDA with *S*-estimators model.

Simulation Study

We applied the classical and robust discriminant rules to the same setting employed in many related research works as shown below (He and Fung, 2000; Croux and Dehon, 2001; Todorov and Pires, 2007):

A. π_1: 50 N_3 (0, I)
   π_2: 50 N_3 (1, I)
B. π_1: 40 N_3 (0, I) +10 N_3 (5, 0.25^2I)
   π_2: 40 N_3 (1, I) +10 N_3 (-4, 0.25^2I)
C. π_1: 80 N_3 (0, I) +20 N_3 (5, 0.25^2I)
   π_2: 8 N_3 (1, I) +2 N_3 (-4, 0.25^2I)
D. π_1: 16 N_3 (0, I) +4 N_3 (0, 25I)
   π_2: 16 N_3 (1, I) +4 N_3 (1, 25I)
E. π_1: 58 N_3 (0, I) +12 N_3 (5, 0.25^2I)
   π_2: 25 N_3 (1, 4I) +5 N_3 (-10, 0.25^2I)
F. π_1: 40 N_3 (0, I) +10 N_3 (5, 25I)
   π_2: 40 N_3 (1, I) +10 N_3 (-4, 25I)

The data was simulated under various conditions that could possibly be encountered in real life. To create these conditions, a few variables were manipulated. These variables were percentage of contamination (17% and 20%); sample sizes (10, 20, 30, 50, 70 and 100); nature of variances (equal and unequal); shift in shape (0.252 and 25) and location (±5).

Condition A was generated from uncontaminated populations while conditions B, C, D, E and F were generated from contaminated populations. 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 about 2000 times for each condition.

In this study, the percentage of contamination and dimension of variables were fixed at 20% and 3, respectively, for conditions A, B, C, D and F. Shift in location with equal and unequal sample sizes were considered in conditions B and C respectively. For condition D, the shift in shape was matched with equal sample sizes. Unequal sample sizes and heteroscedasticity are considered in condition E with almost 17% on contamination percentage. Lastly, extreme contamination was considered in condition F with both location and shape were shifted. Table 1 presents the results of misclassification rate for the classical LDA and RLDA.

From Table 1 we notice that all the models perform equally well when there is no contamination. Theoretically, under ideal condition, that is when all the assumptions are fulfilled, classical LDA should perform optimally and the results in A concur with the theory. Nevertheless, all the RLDA do not perform much worse than the classical LDA. In contrast, when there is contamination, the results show that the misclassification rate for the classical LDA inflated above all the other models (RLDA). In cases B, C and E, the RLDA_M and RLDA_WM perform better than others. They also perform as good as RLDA_S for the rest of the cases (D and F). Furthermore, the proposed models (RLDA_M and RLDA_WM) are more efficient in computational aspect.

Table 1. Mean, standard deviation and computational time of the misclassification rate for various LDA models

		Classical LDA	RLDA_S	RLDA_M	RLDA_WM
A	Mean	0.2001	0.2005	0.2033	0.2007
	s	0.0089	0.0091	0.0108	0.0092
	Time (sec)	3	1177	9	4
B	Mean	0.6512	0.6296	0.2491	0.4104
	s	0.0600	0.0582	0.0416	0.0829
	Time (sec)	3	1176	9	4
C	Mean	0.5004	0.5005	0.4949	0.4984
	s	0.0012	0.0014	0.0082	0.0049
	Time (sec)	3	1223	8	4
D	Mean	0.4442	0.2172	0.2209	0.2237
	s	0.1231	0.0230	0.0241	0.0276
	Time (sec)	3	1156	8	4
E	Mean	0.5007	0.5008	0.4982	0.5002
	s	0.0018	0.0021	0.0048	0.0027
	Time (sec)	3	1114	9	4
F	Mean	0.5814	0.2021	0.2052	0.2127
	s	0.1284	0.0099	0.0119	0.0188
	Time (sec)	3	1101	8	4

Table 2. Results of the lilliefor normality test

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

Table 3. Error rate for the classical LDA and RLDA

	AER	CV
Classical LDA	0.1111	0.1111
RLDA$_S$	0.0741	0.1111
RLDA$_M$	0.0370	0.0741
RLDA$_{WM}$	0.0370	0.0741

Based on Table 1, no one single model can be the best across all the conditions, but taking into account the consistencies of the means and standard deviations of the misclassification rates (which are always small), RLDA$_M$ is the better one. It is comparable to classical LDA under perfect condition (no contamination) and consistently produces small misclassification rate even under contamination of data. The existing RLDA with *S*-estimators performs poorly under a few cases, namely B, C and E.

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.

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 error rates using cross Validation (CV). The results of the real data analysis are presented in Table 3.

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 two proposed models produce the smallest error rates as compared to the existing RLDA$_S$. Both models are found to be equally good as they produce equal error rates via AER as well as CV. The simulation and real life problem results proven that the proposed RLDA models provide a comparable performance or better in LDA.

Conclusion

In this study, we present two robust estimators namely modified one-step *M*-estimator (MOM) and winsorized one-step *M*-estimator (WMOM) to alleviate the classification problem. These two robust estimators used trimming and winsorizing approach to eliminate the outliers of the data and then form the robust discriminant rule. Their function as the substitutes for the classical estimators in Linear Discriminant Analysis (LDA) model 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 two proposed models are comparable or better. The proposed models produce lowest error rates as compared to the other investigated models. Generally, we can conclude that MOM and WMOM estimators should be considered in solving classification problems especially when non-normality and/or heteroscedasticity are suspected. Thus, these new robust models are good alternatives for parametric LDA especially under violation of assumptions.

Acknowledgement

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, Malaysia.

Author's Contributions

The authors contributed significantly in designing the research, analyzing and interpreting the results, drafting and revising the manuscript.

Ethics

This article is original and contains unpublished material. There are no personal conflicts of interest or any ethical issues involved in any aspect of this article.

References

Alrawashdeh, M.J., S.R. Muhammad Sabri and M.T. Ismail, 2012. Robust linear discriminant analysis with financial ratios in special interval. Applied Math. Sci., 6: 6021-6034.

Chorl, C.Y. and P.J. Rousseeuw, 1992. Integrating a high-breakdown option into discriminant analysis in exploration geochemistry. J. Geochem. Explorat., 43: 191-203. DOI: 10.1016/0375-6742(92)90105-H

Croux, C. and C. Dehon, 2001. Robust linear discriminant analysis using s-estimators. Can. J. Stat., 29: 473-493. DOI: 10.2307/3316042

Croux, C., P. Filzmoser and K. Joossen, 2008. Classification efficiencies for robust linear discriminant analysis. Stat. Sinica, 18: 581-599.

Elizabeth, A.M. and M.A. Andres, 2012. Discriminant analysis of multivariate time series using wavelets. Working Paper of Statistics and Economies Series.

GlèlèKakaï, R.M., D. Pelz and R. Palm, 2010. The efficiency of the linear classification rule in multi-group discriminant analysis. Afr. J. Math. Comput. Sci. Res., 3: 019-025.

Haddad, F. S., S.S. Syed-Yahaya and J.L. Alfaro. 2013. Alternative hotelling's T2 charts using winsorized modified one-step m-estimator. Quality Reliability Eng. Int., 29: 583-593. DOI: 10.1002/qre.1407

He, X. and W.K. Fung, 2000. High breakdown estimation for multiple populations with applications to discriminant analysis. J. Multivariate Anal., 72: 151-162. DOI: 10.1006/jmva.1999.1857

Hubert, M. and K. Driessen, 2004. Fast and robust discriminant analysis. Comput. Stat. Data Anal., 45: 301-320. DOI: 10.1016/S0167-9473(02)00299-2

Johnson, R.A. and D.W. Wichern, 2002. Applied Multivariate Statistical Analysis. 5th Edn., Prentice Hall International Edition, New Jersey, ISBN-10: 0130925535, pp: 767.

Lim, Y.F., S.S. Syed Yahaya, F. Idris, H. Ali and Z. Omar, 2014. Robust linear discriminant models to solve financial crisis in banking sectors. Proceedings of the 3rd International Conference on Quantitative Sciences and Its Applications, Aug. 12- 14, AIP, Kedah, pp: 794-798. DOI: 10.1063/1.4903673

Sajobi, T.T., L.M. Lix, B.M. Dansu, W. Laverty and L. Li, 2012. Robust descriptive discriminant analysis for repeated measures data. Comput. Stat. Data Anal., 56: 2782-2794. DOI: 10.1016/j.csda.2012.02.029

Todorov, V. and A.M. Pires, 2007. Comparative Performance of Several Robust Discriminant Analysis Methods. Revstat. Stat. J., 5: 63-83.

Wilcox, R.R. and H.J. Keselman, 2003. Repeated measures one-way ANOVA based on a modified one-step M-estimator. J. Brit. Math. Stat. Psychol., 56: 15-26. DOI: 10.1348/000711003321645313