ROBUST CLASSIFICATION USING MINIMUM REGULARIZED COVARIANCE DETERMINANT ESTIMATOR

R Muthukrishnan and Surabhi S Nair
•
Department of Statistics, Bharathiar University, Coimbatore 641046, Tamil Nadu, India
muthukrishnan1970@gmail.com, surabhinair93@gmail.com

Abstract

The association between a categorical variable and a group of interconnected factors is the main objective of the classification procedure. The linear discriminant analysis (LDA) aims to provide a method for classifying populations and dividing up forthcoming observations among the groups that have already been identified. Under the suppositions of normality and homoscedasticity, the LDA produces the best discriminant rule for two or more groups. Outliers have a significant impact on the parameters of the LDA, mean, and covariance matrix. Robust methods are resistant to outliers. This paper explores the robust methods, namely the Minimum Covariance Determinant (MCD) estimator and Minimum Regularized Covariance Determinant (MRCD) estimators in the context of discriminant analysis under real environments. The MCD technique is used to estimate the location and dispersion matrix using the subset of the given size that has the lowest sample covariance determinant. Its fundamental problem is that it doesn't provide a reliable result when the features/dimension is greater than the size of the subset. As a result, the MRCD method is employed and the efficiency is studied by computing the Apparent Error Rate (AER). In this paper, an attempt has been made to review the existing theory and methods of RLDA.

Keywords: classification, linear discriminant analysis, robust linear discriminant analysis, minimum covariance determinant estimator, minimum regularized covariance determinant estimators

I. Introduction

The traditional Linear Discriminant Analysis (LDA) is a frequently used multi-dimensional classification approach to classify new observations according to one of the aforementioned categories, Elizabeth and Andres [4]. The population means (μ_1, μ_2) with homoscedasticity assumptions $(\Sigma_1 = \Sigma_2 = \Sigma)$ is used for establishing traditional LDA. The Traditional Linear Discriminant Rule will be built using predicted mean vector and scatter matrices while real population parameters are typically unachievable. In particular, the traditional mean is very susceptible to anomalies. Merely just one anomaly can affect the accuracy of the covariance and alter the location estimation, Erceg-hurn et al., [5]. Thus, an immense misclassification rate will be caused by the affected mean and covariances Sajobi et al., [15].

Researchers are looking for solutions in Robust Linear Discriminant Analysis (RLDA) to address this sensitivity issue in LDA. Also can build robust discriminant models with low classification error rates by replacing the classical estimators with robust estimators such as M-estimators, Minimum Covariance Determinant (MCD) estimators, Hubert and Driessen [9], Minimum Volume Ellipsoid (MVE) estimators, Choral and Rousseeuw [2], and S-estimators, He and Fung [8], Croux and Dehon [3], Minimum Regularized Covariance Determinants (MRCD), Boudt and Peter Rousseeuw [1].

This paper mainly compares the robust estimators such as MCD and MRCD in RLDA with traditional LDA. The Apparent Error Rate (AER) is used to calculate how effective certain strategies are. The rest of this paper is structured as follows. Section 2 describes traditional LDA and robust linear discriminant analysis based on MCD and MRCD estimators. The results and discussions based on the real data study will be given in section 3. The conclusion will be provided in the last section.

II. Classification Methods

The traditional classification method, namely, Linear Discriminant Analysis (LDA)), the Robust Linear Discriminant Analysis using Minimum Covariance Determinant (MCD) estimator, and Minimum Regularized Covariance Determinant (MRCD) estimators are briefly discussed in this section.

Linear Discriminant Analysis (LDA)

Fisher [6] introduced the linear discriminant for two classes and C.R. Rao [11] later generalized it for many classes. Linear discriminant analysis (LDA) is a group of multivariate statistical techniques used to identify a linear combination of features that characterize or distinguish two or more classes of objects or events. Hastie et. al. [7], Sharipah et. al. [17]. Classification using a linear function is known as discriminant analysis. The discriminant analysis aims to divide the sample variables into two or more categories. This is accomplished with the use of a linear combination of explanatory factors or forecasting variables. Choosing a group for an object is based on the fundamental tenet that there should be as little chance of misclassification as possible.

On account of the two-group discriminant model, grouping an entity into one of two groups, g_1 or g_2, is the main objective. It is speculated that the explanatory variables will exhibit a multi-variate normal distribution.

$$f(y_1, y_2, \dots, y_p /x = i) = N(Y; \mu, \Sigma)), i = 1,2, \dots, p \quad (1)$$

Let m_i be the number of observations where $x = i, i = 1,2$ and (y_i, x_i) are selected at random using sampling. Alternately, X might be fixed in a way that m_i inspections would be sampled for $X = i$. For each category, sample statistics are computed. The estimates for the sample men $\overline{Y_i}$ and the sample covariance matrix S_i are μ and $\Sigma_i, i = 1,2$ respectively.

Let,

$$V = b_1Y_1 + b_2Y_2 + \dots + b_hY_h$$

$$S = \left(\frac{(m_1m_2)}{m_1+m_2}\right)(\overline{Y_1} - \overline{Y_2})(\overline{Y_1} - \overline{Y_2})'$$

The pooled within the covariance matrix is

$$U_b = \left(\frac{(m_1m_2)}{m_1+m_2}\right)\left\{\frac{b'(\overline{Y_1}-\overline{Y_2})(\overline{Y_1}-\overline{Y_2})'b}{b'Sb}\right\} \quad (2)$$

This is the generalized eigenvalue problem given by

$$Ab = cSb;\ A = \left(\frac{(m_1m_2)}{m_1 + m_2}\right)(\overline{Y_1} - \overline{Y_2})(\overline{Y_1} - \overline{Y_2})'$$

The clarification b is proportional to $S^{-1}(\overline{Y_1} - \overline{Y_2})$. so the sample discriminant variable, $V = (\overline{Y_1} - \overline{Y_2})S^{-1}Y$. It is the linear combination of the original observation which has the largest ratio of the between-groups to the within-group variation.

When X has p categories, then p groups (g_1, g_2… g_p) have been established for the p group situations. Let m_i be the number of clarifications in the ith group, $m = \sum_i^p m_i$. The sample mean and covariance matrix is given by $\overline{Y_l}$ and S_i. The matrix U represents the pooled within-group covariance.

$$U = \frac{1}{(m-p)} \sum_{i=1}^{p} (m_i - 1) S_i.$$

The between-group covariance matrix is

$$A = \frac{1}{p-1} \sum_i^p m_i \, (\overline{Y_l} - \overline{Y})(\overline{Y_l} - \overline{Y})' \tag{3}$$

The traditional method relies heavily on the sample mean vector and covariance matrix, which are vulnerable to out-of-the-ordinary data. Additionally, when used to datasets with smashed model assumptions, the LDA model may yield erroneous outcomes. A robust approach can be used to solve this problem.

Robust Linear Discriminant Analysis (RLDA)

RLDA has been developed as a modified version of traditional LDA, especially for handling non-ideal data such as outliers, high dimensions, and multicollinearity, Sharipah et. al. [17]. In robust methods, classical mean vectors and covariance matrices are replaced by robust counterparts, Peter J. Rousseeuw, Mia Hubert [14], and Muthukrishnan et. al. [10]. The robust estimators used here in the robust linear discriminant analysis are MCD and MRCD.

Minimum Covariance Determinant Estimator (MCD)

Rousseeuw [12] developed the Minimum Covariance Determinant (MCD) Estimator to estimate the mean vector and covariance matrix as well as to identify outliers. The lowest determinant-containing subset of h observations with respect to their covariance matrix is sought after. The location estimate is the mean value of that subgroup according to this estimator, and the scatter estimate is a multiple of its scatter matrix.

$$M_X(H) = h^{-1} X_H' I_h \tag{4}$$

$$S_X(H) = (h-1)^{-1} (X_H - M_X(H))' (X_H - M_X(H)) \tag{5}$$

After that, the MCD method seeks to minimize the determinant of $S_X(H)$ for all $H \in \mathcal{H}_h$.

$$H_{MCD} = \underset{H \in \mathcal{H}_h}{argmin} \left(det(S_X(H))^{1/p} \right) \tag{6}$$

For statistical considerations, eqn (3) takes the p^{th} root of the determinant. The geometric average of its eigenvalues is the p^{th} root of the determinant of the scatter matrix. It is referred to as the standardized generalized variance by Sen Gupta (1987).

The mean of the h-subset is used to define the MCD estimate of location M_{MCD}, while the MCD scatter estimate is expressed as a multiple of the sample scatter matrix, and is given by

$$M_{MCD} = M_X(H_{MCD}) \tag{7}$$
$$S_{MCD} = C_\alpha S_X(H_{MCD}) \tag{8}$$

where C_α is a consistency factor that is based on the trimming percentage = $(n-h)/n$ and is similar to the one provided by Croux and Haesbroeck [3]. Its fundamental flaw is that it gives unreliable results when the dimension is greater than the size of the subset. In high dimensions, it is necessary to modify MCD, as the existing MCD methods are slow and less robust in that situation.

Minimum Regularized Covariance Determinant Estimator (MRCD)

The MRCD estimator was proposed by Boudt et. al. [1]. To guarantee that the MRCD scatter estimator is scale equivariant and location unvarying, as is common in the literature, first, standardize the variables. The use of a trustworthy univariate location and scale estimate is required for standardization. For this, the median of each subset is calculated and placed in a location vector called m_x. Additionally, each variable's scale using the Qn estimator of Rousseeuw and Croux (1993) is calculated, then insert these scales into d_x, the diagonal matrix. The standardized observation is given by

$$Z_i = d_x^{-1}(x_i - m_x) \tag{9}$$

The regularized scatter matrix of the standardized observation is

$$\mathcal{S}(H) = \rho T + (1-\rho)C_\alpha S_Z(H)$$

where $S_Z(H)$ is defined in (5), however, in the case of Z, c is the same consistency parameter as in (8).

Let A be the diagonal matrix containing eigenvalues of T, and the orthogonal matrix Q contains the relevant eigenvectors. Utilizing the spectral decomposition $T = QAQ'$ will be practical.

Now,

$$\mathcal{S}(H) = QA^{1/2}[\rho I + (1-\rho)C_\alpha S_W(H)]AA^{1/2}Q' \tag{10}$$

where W is the $n \times p$ matrix consisting the transformed standardized observations

$$w_i = A^{-1/2}Q'Z_i, \text{ and}$$

$$S_W(H) = A^{-1/2}Q'S_Z QA^{-1/2}$$

The MRCD subset is given by

$$H_{MRCD} = \underset{H\in\mathcal{H}_h}{argmin}\left(det(\rho I + (1-\rho)C_\alpha S_W(H))^{1/p}\right) \tag{11}$$

The MRCD location and scatter estimations of the initial data matrix X are defined as follows

$$M_{MRCD} = m_X + d_x M_Z(H_{MRCD})$$

$$S_{MRCD} = C_\alpha S_X(H_{MCD})$$

III. Experimental Results

Table 1: *Apparent Error Rate under Classical and Robust Methods*

Dataset	LDA	RLDA	
		MCD	MRCD
Hemophilia	0.146 (0.150)	0.146 (0.146)	0.133 (0.125)
Anorexia	0.513 (0.528)	0.486 (0.457)	0.388 (0.371)
(.) without outliers			

Table 2. *Classification Matrix of Hemophilia Data under Classical and Robust Methods*

		RLDA	
Methods	LDA	MCD	MRCD
With outliers	$\begin{pmatrix}38 & 7\\ 4 & 26\end{pmatrix}$	$\begin{pmatrix}38 & 7\\ 4 & 26\end{pmatrix}$	$\begin{pmatrix}39 & 6\\ 4 & 26\end{pmatrix}$
Without outliers	$\begin{pmatrix}37 & 7\\ 4 & 25\end{pmatrix}$	$\begin{pmatrix}38 & 7\\ 4 & 26\end{pmatrix}$	$\begin{pmatrix}38 & 76\\ 3 & 26\end{pmatrix}$

Table 3. *Classification Matrix of Anorexia Data under Classical and Robust Methods*

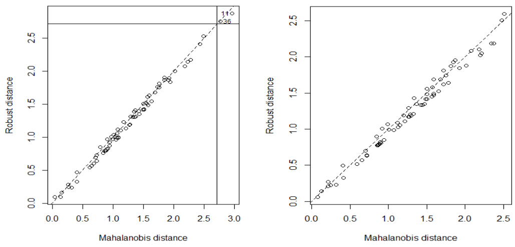
		RLDA	
Methods	LDA	MCD	MRCD
With outlier	$\begin{pmatrix}11 & 10 & 8\\ 9 & 17 & 0\\ 6 & 4 & 7\end{pmatrix}$	$\begin{pmatrix}16 & 5 & 8\\ 16 & 8 & 2\\ 0 & 4 & 13\end{pmatrix}$	$\begin{pmatrix}15 & 5 & 8\\ 0 & 16 & 4\\ 0 & 4 & 13\end{pmatrix}$
Without outlier	$\begin{pmatrix}10 & 2 & 6\\ 9 & 17 & 0\\ 6 & 4 & 6\end{pmatrix}$	$\begin{pmatrix}16 & 5 & 7\\ 14 & 10 & 2\\ 0 & 4 & 12\end{pmatrix}$	$\begin{pmatrix}15 & 6 & 7\\ 5 & 17 & 4\\ 0 & 4 & 12\end{pmatrix}$

This section examined the effectiveness of traditional and robust methods of discriminant analysis techniques in terms of classification problems. For the study, the two actual data sets were taken into account. The first one is the hemophilia data set from the R package named rrcov. There are two assessed factors in the hemophilia data. AHF action and AHV antigen upon 75 women, divided into two groups, namely, compulsory carriers, which includes 45 data points, and the normal group, which includes 30 data points.

The second one is the anorexia data on weight change from the R package named *MASS,* and is divided into three groups, each of which has two variables and a set of 72 occurrences; Information on young female anorexic patients' weight changes. Prewt (patient weight prior to study times) and Postwt (patient weight following study times) are the two variables used to classify the three groups into Cognitive Behavioral Treatment (CBT), Control (Cont), and FT family treatment (FT).

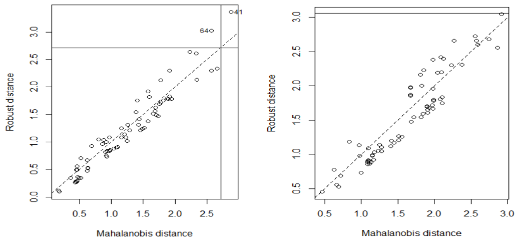
These data sets experienced classification analysis using classic LDA and alternative RLDA algorithms under with/without outliers. Distance-distance plots were used to identify the anomalies (Figure 1, Figure 2).). On the basis of their Classification matrix and Apparent Error Rate (AER), the classification criteria are assessed. The classification matrix is just a table where the rows represent the dependent categories that were observed and the columns represent the expected dependent categories. All examples will fall on the diagonal if the prediction is perfect. The proportion of correctly classified cases is represented by the diagonal cases. The results achieved under various methods are concluded in the form of Apparent Error Rate (Table 1) and classification matrix (Table 2 and Table 3). Robust classification procedure using MRCD estimator gives less Apparent Error Rate and more classification accuracy when compared with other classification procedures.

The result reveals that robust procedures provide better results when compared with the traditional method, Linear Discriminant Analysis. Further, it is observed that MRCD estimator based RLDA outperforms over MCD estimator.



with outliers without outliers

Figure 1*: Distance-Distance Plot (Haemophilia dataset)*



with outliers without outliers

Figure 2: *Distance-Distance Plot (Anorexia dataset)*

IV. Conclusion

Classification analysis is one of the key concepts in the context of statistical learning. This study explores classification analysis using traditional and various approaches of robust linear discriminant analysis. Conventional methods should work reasonably well if certain assumptions are true, however, they may not be trustworthy if one or more of these assumptions are erroneous. Both sample mean vector and covariance matrix are extremely susceptible to anomalies. As a result, when the data contains anomalies, the traditional LDA fails to generate reliable results. For non-normal conditions, a robust alternative is required to improve accuracy even when the data somewhat depart from the model assumptions. When robust estimators such as MCD and MRCD are used in LDA, the analysis performs well compared to traditional methods. Robust methods perform well even when the model assumptions are not met. The study came to the conclusion that robust classification using MRCD estimators offers greater accuracy, followed by other methods. The study could be expanded by using appropriate robust estimators in RLDA to further improve accuracy.

References

[1] Boudt K., Rousseeuw P.J. and Vanduffel S. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30:113–128.

[2] Choral C. Y. and Rousseeuw P. J. (1992). Integrating a high-breakdown option into discriminant analysis in exploration geochemistry, *Journal of Geochemical Exploration*, 43:191-203.

[3] Croux C. and Dehon C. (2001). Robust linear discriminant analysis using S-estimators, *Canadian Journal of Statistics*, 29:473-493.

[4] Elizabeth A. M. and Andres M. A. (2016). Discriminant analysis of multivariate time series: Application to diagnosis based on ECG signals. *Computational Statistics and Data Analysis*, *70*:67 – 87.

[5] Erceg-Hurn D. M., Wilcox R. R. and Keselman H. J. (2013). Robust statistical estimation. *In T. D. Little (Ed). The Oxford Handbook of Quantitative Methods: Foundations, Oxford University Press*, 1: 388-406.

[6] Fisher R. A. (1936). The use of multiple Measurements in Taxonomic problem. *The Annals of Eugenics*, 7:179-188.

[7] Hastie T. Tibshirani R. and Friedman J. H. (2009). Elements of statistical learning. *Springer, New York.*

[8] He X. and Fung W. K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, 72:151-162.

[9] Hubert M. and Driessen K. (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45:301-320.

[10] Muthukrishnan R. and Udaya Prakash N. (2019). Performance of Classification Techniques along with Support Vector Machine. *International Journal of Innovative Technology and Exploring Engineering*, 9:2278-3075.

[11] Rao C. R. (1948). Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation. *Proceedings of the Cambridge Philosophical Society*, 44:50-57.

[12] Rousseeuw P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

[13] Rousseeuw P. J. Croux C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283.

[14] Rousseeuw P. J. and Mia Hubert (2017). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*, .8, 10.1002/widm.1236.

[15] Sajobi T. T., Lix L. M., Dansu B. M., Laverty W. and Lix L. (2012). Discriminant Analysis for Repeated Measures Data: Effects of Mean and Covariance Misspecification on Bias and Error in Discriminant Function Coefficients. *Journal of Modern Applied Statistical Methods*, 56:2782-2794.

[16] SenGupta A. (1987). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *Journal of Multivariate Analysis*, 23:209–219.

[17] Sharipah Soaad S. Y., Lim Y. F., Hazlina Ali and Zurni Omar. (2016). Robust Linear Discriminant Analysis. *Journal of Mathematics and Statistics,* 12:312-316.

[18] Todorov V. (2007). Robust selection of variables in linear discriminant analysis. *Statistical Methods and Applications*, 5:395-407.