Journal of Engineering and Applied Sciences 13 (24): 10396-10400, 2018
ISSN: 1816-949X

Modified One-Step M-Estimator with Robust Scale Estimator for Multivariate Data

Hameedah Naeem Melik, Nor Aishah Ahad and Sharipah Soaad Syed Yahaya
School of Quantitative Sciences, Universiti Utara Malaysia, Sintok, Kedah, Malaysia

Abstract: The Modified One-step M-estimator (MOM) is a highly efficient robust estimator for classifying multivariate data. Generally, robust estimators came into existence as a solution to the inability of classical Linear Discriminant Analysis (LDA) to perform optimally in the presence of outliers. Thus, to solve this shortcoming, the robust MOM estimator is integrated with a highly robust scale estimator, Q_n, in the trimming criterion of MOM. This introduces a new robust approach termed $RLDA_{MQ}$ for handling outliers encountered in multivariate data. The results show the superiority of $RLDA_{MQ}$ over the classical LDA and previously existing robust method in literature in terms of misclassification error evaluated through simulated data.

Key words: Modified one-step M-estimator, robust, Q_n, multivariate data, trimming criterion, encountered

INTRODUCTION

Discriminant analysis is a statistical classification technique where the object groups and several training examples of objects that have been grouped are known and the model of classification is also given. Discriminant analysis is one of the methods that give more information to the structure of multivariate data which are data arising from variables greater than one (Fidler and Leonardis, 2003; Cacoullos, 2014). Certain features of a discriminant analysis platform include the choice of fitting methods which births the common discriminant analysis method; Linear Discriminant Analysis (LDA).

LDA as introduced by Fisher (1936) is a very imperative and archetypal technique in discriminant analysis as it has good use in practical applications. LDA performs well for data that follow normal distribution with identical population covariance matrices but shows instability when the assumptions are violated (Omidiora *et al*., 2008; Croux *et al*., 2008). Discriminant analysis has a high level of vulnerability to outliers which is seen to be present in many real world multivariate data sets (Sajtos and Mitev, 2007). This is likewise the case when considering LDA which due to the constraint that the LDA parameters are highly affected by outlying observations gives room for misclassification of new observations (Kim *et al*., 2006; Pires and Branco, 2010; Jin and An, 2011). These setbacks caused researchers to venture into introduction of robust estimators which will most importantly handle the presence of outliers (Cheng *et al*., 2016).

A wide range of robust estimators exist and have been adopted for handling the outliers in the data that the conventional LDA approach cannot handle (Filzmoser and Todorov, 2013; Todorov and Pires, 2007), ranging from the M-estimators, Minimum Volume Ellipsoid (MVE), Minimum Covariant Determinant (MCD) and S-estimators as introduced by Campbell (1980), Rousseeuw (1984, 1985) and Davies (1987), respectively. The concept for robustifying LDA involves the replacement of the classical mean vectors and covariance matrices by its robust counterparts. This approach known as the plug-in method has been adopted in various way to introduce new robust linear discriminant analysis methods (Sajobi *et al*., 2012; Alrawashdeh *et al*., 2012; Todorov and Pires, 2007; Filzmoser *et al*., 2006).

Some specific literature includes the researches of Yahaya *et al*. (2016) and Lim *et al*. (2016) who introduced an automatic trimmed mean vector and a winsorized approach as a substitute for the classical mean vector. On the other hand, a robust approach of multiplying the Spearman's rho with the corresponding robust scale estimator was adopted by these researchers as a substitute for the covariance matrices. Similarly, this article will be adopting the plug-in method to introduce a new robust linear discriminant analysis method. The robust estimator adopted to replace the classical mean vector is the Modified One-step M-estimator (MOM) by Wilcox and Keselman (2003) integrated with a highly robust scale estimator Q_n in the trimming criteria. This new robust estimator introduced in this study is coined $RLDA_{MQ}$.

MATERIALS AND METHODS

This study presents a brief description of the conventional LDA algorithm and the new robust approach $RLDA_{MQ}$.

Corresponding Author: Hameedah Naeem Melik, School of Quantitative Sciences, Universiti Utara Malaysia, Sintok, Kedah, Malaysia

LDA: Consider a two-group discrimination problem with n observations of a training data measured at d characteristics are given. The n observations are obtained from two different populations, π_1 and π_2 with corresponding sample sizes, n_1 and n_2. The classical LDA rule as defined by Johnson and Wichern (2014) is given in Eq. 1:

$$\text{If:} \quad (\mu_1 - \mu_2)^t \sum^{-1} \left[x_0 - \frac{1}{2}(\mu_1 - \mu_2) \right] \geq \ln\left(\frac{p_2}{p_1} \right) \quad (1)$$

then: $x_0 \in \pi_1$

otherwise: $x_0 \in \pi_1$

Where:

p_1 = The prior probability that an individual comes from population π_1

p_2 = The prior probability that an individual comes from population π_2

Note that the classification of the observation x_0 will be optimal only if the assumption that π_1 and π_2 are both multivariate normal distributions with different location but having identical covariance is satisfied (Lim *et al*., 2016). In addition, if there are outliers in the training data, then the estimators of mean and covariancecan be seriously affected. Thus, this brings the introduction of the new robust method as seen in the next subsection.

$RLDA_{MQ}$: This approach involves combining the MOM statistic with the highly robust Q_n scale estimator. MOM as obtained from the conventional one-step M-estimator (Haddad, 2013; Staudte and Sheather, 1990) but with certain modifications is simply the average of the values remaining after the removal of all extreme values (if there is existence of any). The robust Q_n scale estimator on the other hand as proposed by Rousseeuw and Croux (1993) is a well suitable estimator with the advantage of high efficiency. The algorithm to combine these two techniques is described iteratively as:

LDA algorithm:

Step 1: Trim the data to be analyzed using the default scale estimator MAD_n for determining the extreme values in MOM criterions. Let $\hat{M}_j$ be the median for group j:

$$MAD_{nj} = \frac{MAD_j}{0.6745};\ MAD_j = \text{Median}\left|Y_{1j} - \hat{M}_j\right|, \left|Y_{2j} - \hat{M}_j\right|, ..., \left|Y_{nj} - \hat{M}_j\right|$$

Step 2: Compute $\hat{\theta}$ from:

$$\hat{\theta}_j = \frac{\sum_{i=i_1+1}^{n_j - i_2} Y_{(i)j}}{n_j - i_1 - i_2}$$

Step 3: Calculate Q_n from $Q_n = 2.2219\{|X_i\text{-}X_j|;\ i<j;\ 1, 2, 3, ..., n;\ j = 1, 2, 3, ..., n\}_{(k)}$

Step 4: Replace the default scale estimator MAD_n in step 2 with the Q_n estimator to obtain i_1 as the number of observations Y_j such that $(Y_{ij} - \hat{M}_j) < -2.24(Q_{nj})$ and i_2 is the number of observations $(Y_{ij} - \hat{M}_j) > -2.24(Q_{nj})$

Step 5: Compute $\hat{\theta}_j$ based on the Q_n estimator in step 4.

The next study will consider the implementation of the new robust method with simulated data. Comparison is made with the classical LDA and other robust approach in literature.

RESULTS AND DISCUSSION

The performance of the methods computed in terms of misclassification error was investigated on several simulation conditions involving manipulation of five variables as shown in Table 1. The choice of variables to be manipulated follows from prior adoption in previous studies such as Haddad (2013) and Lim *et al*. (2016) amongst others.

The combination of various variable settings produced 306 different data distributions (18 uncontaminated, 72 location contamination, 72 shape contamination and 144 location and shape contamination). Each group π_j, j = 1, 2 has a separate mean μ_j but the same covariance matrix I_p. Therefore, the data was contaminated for the covariance matrices as follows:

$$\begin{aligned} \pi_1 &: (1\text{-}\varepsilon)N_p(\mu_j, I_p) + \varepsilon N_p(\mu_j + \mu, \kappa I_p) \\ \pi_2 &: (1\text{-}\varepsilon)N_p(\mu_j, I_p) + \varepsilon N_p(\mu_j - \mu, \kappa I_p) \end{aligned} \quad (2)$$

A testing sample of size 2000 from each population was generated and the misclassification error was computed by obtaining the proportion of misclassified testing sample observations in each population. The simulation process was repeated 2000 times and the mean

Table 1: Simulation conditions

Variable	Descriptions
Dimension of variable (d)	2, 6
Percentage of contamination (ε)	0, 10, 20
Sample size of the training data (n_1, n_2)	(20, 20), (50, 50), (100, 100)
Shift in location of the population (μ)	0, 3, 5
Shift in shape of the population (κ)	0, 9, 25

Table 2: Mean misclassification error for linear discriminant models with

			$(n_1, n_2) = (20, 20)$			$(n_1, n_2) = (50, 50)$			$(n_1, n_2) = (100, 100)$		
ε	μ	κ	LDA	Lim *et al.* (2016)	$RLDA_{MQ}$	LDA	Lim *et al.* (2016)	$RLDA_{MQ}$	LDA	Lim *et al.* (2016)	$RLDA_{MQ}$
-	-	-	**0.2511**	0.2543	0.2530	**0.2442**	0.2548	0.2449	**0.2420**	0.2429	0.2424
10	3	-	0.3389	**0.2866**	0.2867	0.2960	0.2646	**0.2583**	0.2741	0.2542	**0.2496**
10	5	-	0.4987	0.2862	**0.2723**	0.4986	0.2658	**0.2519**	0.5010	0.2566	**0.2462**
10	0	9	0.3178	0.2579	**0.2549**	0.2759	0.2472	**0.2455**	0.2587	0.2438	**0.2427**
10	0	25	0.4205	0.2579	**0.2542**	0.3863	0.2474	**0.2452**	0.3447	0.2439	**0.2426**
10	3	9	0.3884	0.2602	**0.2556**	0.3610	0.2487	**0.2456**	0.3270	0.2446	**0.2428**
10	3	25	0.4527	0.2587	**0.2544**	0.4441	0.2479	**0.2453**	0.4234	0.2441	**0.2426**
10	5	9	0.4548	0.2631	**0.2570**	0.4732	0.2502	**0.2461**	0.4804	0.2455	**0.2430**
10	5	25	0.4755	0.2593	**0.2545**	0.4870	0.2483	**0.2452**	0.4917	0.2444	**0.2426**
20	3	-	0.5770	0.4753	**0.4745**	0.6202	0.5297	**0.4009**	0.6542	0.5772	**0.3480**
20	5	-	0.6530	0.4442	**0.3925**	0.6911	0.5179	**0.2998**	0.7124	0.6010	**0.2710**
20	0	9	0.3624	0.2628	**0.2608**	0.3055	0.2499	**0.2470**	0.2745	0.2451	**0.2433**
20	0	25	0.4637	0.2622	**0.2576**	0.4277	0.2499	**0.2461**	0.3929	0.2454	**0.2429**
20	3	9	0.5083	0.2735	**0.2624**	0.5334	0.2561	**0.2479**	0.5678	0.2489	**0.2437**
20	3	25	0.5041	0.2652	**0.2574**	0.5062	0.2515	**0.2463**	0.5237	0.2461	**0.2430**
20	5	9	0.6039	0.2865	**0.2662**	0.6795	0.2665	**0.2492**	0.7158	0.2565	**0.2445**
20	5	25	0.5310	0.2678	**0.2578**	0.5590	0.2530	**0.2465**	0.6061	0.2469	**0.2431**

Table 3: Mean misclassification error for linear discriminant models with

			$(n_1, n_2) = (20, 20)$			$(n_1, n_2) = (50, 50)$			$(n_1, n_2) = (100, 100)$		
ε	μ	κ	LDA	Lim *et al.* (2016)	$RLDA_{MQ}$	LDA	Lim *et al.* (2016)	$RLDA_{MQ}$	LDA	Lim *et al.* (2016)	$RLDA_{MQ}$
-	-	-	**0.1409**	0.1481	0.1442	**0.1214**	0.1246	0.1226	**0.1157**	0.1173	0.1163
10	3	-	0.3915	0.2733	**0.2728**	0.3286	0.2123	**0.1937**	0.2740	0.1759	**0.1574**
10	5	-	0.4998	0.2758	**0.2438**	0.5004	0.2184	**0.1697**	0.4991	0.1855	**0.1418**
10	0	9	0.2108	0.1529	**0.1484**	0.1812	0.1276	**0.1247**	0.1505	0.1189	**0.1172**
10	0	25	0.2543	0.1535	**0.1481**	0.2696	0.1280	**0.1246**	0.2252	0.1192	**0.1172**
10	3	9	0.2679	0.1631	**0.1541**	0.2757	0.1338	**0.1271**	0.2414	0.1224	**0.1184**
10	3	25	0.2655	0.1557	**0.1485**	0.3288	0.1298	**0.1250**	0.3142	0.1201	**0.1173**
10	5	9	0.3253	0.1754	**0.1625**	0.3809	0.1412	**0.1306**	0.4000	0.1267	**0.1202**
10	5	25	0.2783	0.1581	**0.1497**	0.3812	0.1313	**0.1255**	0.4072	0.1210	**0.1175**
20	3	-	0.5365	0.4659	0.4698	0.5611	0.5070	**0.4313**	0.5866	0.5399	**0.3913**
20	5	-	0.5668	0.4436	**0.4141**	0.6101	0.4896	**0.3300**	0.6526	0.5459	**0.2670**
20	0	9	0.2514	0.1603	**0.1567**	0.1980	0.1321	**0.1277**	0.1587	0.1212	**0.1185**
20	0	25	0.3613	0.1607	**0.1541**	0.3534	0.1327	**0.1270**	0.2921	0.1218	**0.1181**
20	3	9	0.3933	0.1842	**0.1693**	0.4948	0.1507	**0.1338**	0.5381	0.1330	**0.1214**
20	3	25	0.4204	0.1657	**0.1553**	0.4977	0.1366	**0.1277**	0.5044	0.1242	**0.1185**
20	5	9	0.4956	0.2167	**0.1882**	0.6776	0.1805	**0.1433**	0.7669	0.1554	**0.1265**
20	5	25	0.4625	0.1711	**0.1572**	0.5911	0.1407	**0.1287**	0.6490	0.1266	**0.1190**

Bold values are significant

misclassification error was recorded as seen in Table 2 and 3. The performance percentage of each model, that is, the model with the least misclassification error is also displayed graphically in Fig. 1 and 2.

Considering the percentage of contamination (ε), Table 1 states that the percentage will vary from 10-20%. It is observed that as increases, the mean misclassification error also increases at constant μ and κ. This confirms that the presence of contamination in data makes it difficult for linear models to correctly classify this data. Although, the robust models perform considerably better than the classical LDA with increased contamination. Thus, considering the performance of each linear model, Fig. 1 and 2 show the performance percentages. From both Fig. 1 and 2 and Table 1-3, LDA obtains the least mean misclassification error when the data is clean (no contamination) with ε = 0, μ = 0, κ = 0. Although, as soon as there is contamination in the data, the better performance shifted to the robust models. Comparison is

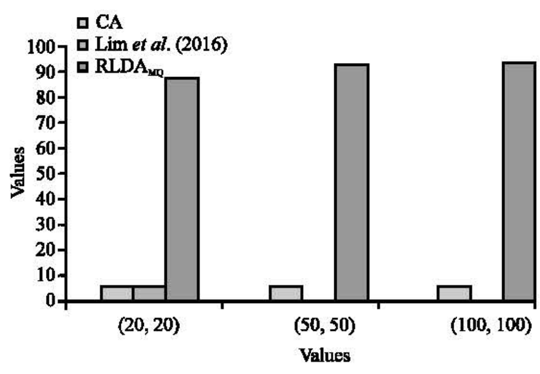

Fig. 1: Performance percentage chart for linear discriminant models with d = 2

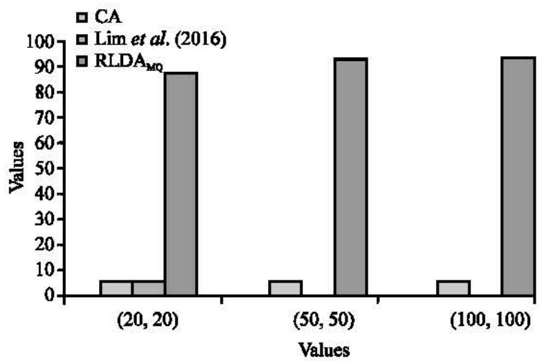

Fig. 2: Performance percentage chart for linear discriminant models with d = 6

made between the robust model presented by Lim *et al.* (2016) and the new RLDA$_{MQ}$ Model. It is observed that for both dimensions 2 and 6, RLDA$_{MQ}$ has the highest performance percentage giving more impressive results that the results from Lim *et al.* (2016). Moving over to the shift in location of the population, one notable behavior as increased from 3-5 at is the decrease in the mean misclassification error for the robust models. Whereas, for the shift in shape of the population, the general behavior for all the linear discriminant models is sharp reduction as soon as goes from 0-9 before its gradual descent to convergence.

Therefore, a general overview of the results obtained in Table 1 and 2 show CA obtaining the least mean misclassification error at no contamination. This is in line with the theory that the classical LDA approach will perform optimally when the assumptions of the LDA are fulfilled. Although, RLDA$_{MQ}$ and Lim *et al.* (2016) also gave favourable results as the difference between the mean misclassification error of the robust estimators and the classical approach is very small which shows convergence in results. However, as soon as there is contamination in the data, the better models are the robust models with RLDA$_{MQ}$ performing better than (Lim *et al.*, 2016).

In addition, it is also observed that the misclassification error is inversely proportional to the dimension of the variables, that is as d increases, mean misclassification error reduces, except when there is no shift in shape of the population ($\kappa = 0$). For instance, when considering the increase from d = 2 to d = 6, the mean misclassification error reduces to about half of its initial value. However, this pattern is not observed for the classical model which did not display such convergence with respect to the increase in the dimension of the variables.

CONCLUSION

This study has presented a new robust approach that is suitable for handling outliers in multivariate data. The robust model considered the modified one-step m-estimator integrated with the Q_n scale estimator. The resulting robust RLDA$_{MQ}$ Model was compared with the classical LDA approach and another robust model proposed by Lim *et al.* (2016) using certain simulated data. It was observed from the mean misclassification error that the RLDA$_{MQ}$ performs that both linear approaches. Therefore, RLDA$_{MQ}$ is a suitable approach to solve the classification problems even under various cases of contamination in data sets.

REFERENCES

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

Cacoullos, T., 2014. Discriminant Analysis and Applications. Elsevier, Amsterdam, Netherlands, ISBN:9781483268712, Pages: 456.

Campbell, N.A., 1980. Robust procedures in multivariate analysis I: Robust covariance estimation. Appl. Stat., 29: 231-237.

Cheng, G., X. Li, P. Lai, F. Song and J. Yu, 2017. Robust rank screening for ultrahigh dimensional discriminant analysis. Stat. Comput., 27: 535-545.

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

Davies, P.L., 1987. Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. Ann. Stat., 15: 1269-1292.

Fidler, S. and A. Leonardis, 2003. Robust LDA classification by subsampling. Proceedings of the Workshop on Computer Vision and Pattern Recognition (CVPRW'03) Vol. 8, June 16-22, 2003, IEEE, Madison, Wisconsin, USA., ISBN:0-7695-1900-8, pp: 97-97.

Filzmoser, P. and V. Todorov, 2013. Robust tools for the imperfect world. Inf. Sci., 245: 4-20.

Filzmoser, P., K. Joossens and C. Croux, 2006. Multiple Group Linear Discriminant Analysis: Robustness and Error Rate. In: Compstat 2006-Proceedings in Computational Statistics, Rizzi, A. and M. Vichi (Eds.). Physica-Verlag, Heidelberg, Germany, pp: 521-532.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen., 7: 179-188.

Haddad, F.S., 2013. Statistical process control using modified robust hotelling's T2 control charts. Ph.D. Thesis, Universiti Utara Malaysia, Changlun, Malaysia.

Jin, J. and J. An, 2011. Robust discriminant analysis and its application to identify protein coding regions of rice genes. Math. Biosci., 232: 96-100.

Johnson, R.A. and D.W. Wichern, 2014. Applied Multivariate Statistical Analysis. Vol. 4, Prentice-Hall, New Jersey, USA., ISBN-13:978- 013187715,.

Kim, S.J., A. Magnani and S. Boyd, 2006. Optimal kernel selection in kernel fisher discriminant analysis. Proceedings of the 23rd International Conference on Machine Learning, June 25-29, 2006, ACM, Pittsburgh, Pennsylvania, USA., ISBN:1-59593-383-2, pp: 465-472.

Lim, Y.F., S.S.S. Yahaya and H. Ali, 2016. Winsorization on linear discriminant analysis. Proceedings of the 4th International Conference on Quantitative Sciences and Its Applications Vol. 1782, August 16-18, 2016, AIP Publishing, Melville, New York, pp: 0500101-0500110.

Pires, A.M. and J.A. Branco, 2010. Projection-pursuit approach to robust linear discriminant analysis. J. Multivariate Anal., 101: 2464-2485.

Rousseeuw, P.J. and C. Croux, 1993. Alternatives to the median absolute deviation. J. Am. Statist. Assoc., 80: 1273-1283.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc., 79: 871-880.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. Math. Statist. Appl., 13: 283-297.

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.

Sajtos, L. and A. Mitev, 2007. SPSS Research and Data Analysis Handbook. Alinea Publisher, Budapest, Hungary,.

Staudte, R.G. and S.J. Sheather, 1990. Robust Estimation and Testing. 2nd Edn., John Wiley and Sons, New York, ISBN: 978-0-471-85547-7.

Todorov, V. and A.M. Pires, 2007. Comparative performance of several robust linear discriminant analysis methods. Revstat Stat. J., 5: 63-83.

Wilcox, R.R. and H.J. Keselman, 2003. Modern robust data analysis methods: measures of central tendency. Psychol. Methods, 8: 254-274.

Yahaya, S.S.S., Y.F. Lim, H. Ali and Z. Omar, 2016. Robust linear discriminant analysis with automatic trimmed mean. J. Telecommun. Electron. Comput. Eng., 8: 1-3.