Robust Linear Discriminant Analysis with Financial Ratios in Special Interval

Mufda Jameel Alrawashdeh
School of Mathematical Sciences, University Sains Malaysia
11800 USM, Penang, Malaysia
mufdajr@yahoo.com

Shamsul Rijal Muhammad Sabri
School of Mathematical Sciences, University Sains Malaysia
11800 USM, Penang, Malaysia

Mohd Tahir Ismail
School of Mathematical Sciences, University Sains Malaysia
11800 USM, Penang, Malaysia

Abstract
The linear discriminant analysis (LDA) is a method that can be utilized in a multi-group setting. The basic idea of LDA is to classify an object of unknown origin to one of several given classes based on the measurement vector and is widely used features extraction method in classification. However the LDA has many applications to investigate the performance and extract the features of companies. The financial ratios had obtained from corporate balance sheets, hence almost of the financial ratios spread in the interval [0, 1]. The paper presents study the performance of robust linear discriminant analysis (RLDA) through real data and simulation study where the data generated based on means and covariance matrices values in between [0,1]. The results show very well performance for RLDA same as a good agreement results at real data and simulation study.

Keywords: Linear Discriminant Analysis, Fast-MCD estimator, Financial Ratios

1 Introduction
The linear discriminant analysis (LDA) is a method that can be utilized in a multi-group setting. The basic idea of LDA is to classify an object of unknown origin to one of several given classes based on the measurement vector
(also called discriminator) within p-dimensional space. Where the availables data sets are samples of objects may be combined so as to enhance one’s understanding of groups differences. Practically the parameters of the LDA are unknown; they have to estimate from the sample data. In the classical approach discriminant analysis rules are often based on the empirical mean and covariance matrix of sample data. Because these parameters are highly affect by outlying observations, robust discriminant analysis rules are proposed to overcome sensitivity problem of discriminant analysis rules. High-breakdown criterion developed for LDA to overcome the outlier’s problem was presented by Hawkins and McLachlan (1997). Rousseeuw and Levoy (1987) studied the S-estimator by applying LDA. For more efficiency the observations are pooled to estimate the common covariance matrix instead of averaging the robust estimates of the individual covariance matrices He and Fung (2000). ”A Fast Algorithm for the Minimum Covariance Determinant Estimator” (Fast-MCD) is highly robust to outliers which proposed by Rousseeuw and Driessen (1999). Fast-MCD was applied by plugging the MCD estimator for each discriminant rules which was taken by Hubert and Driessen (2004). The arguments above showed that robust linear discriminant analysis is performs well against the sensitivity problems and it is computed very fast for large data sets.

In order to evaluate the financial performance of a company the financial ratios are employed as evaluation criterion. The earliest study using discriminant analysis data on failure prediction was conducted by Altman (1968) utilizing set of financial ratios to determinants of corporate failures. The financial ratios as the independent variable were entered into the discriminant analysis function, and the highly correct categorization is achieved, Sori and Jalil (2009). Abdullah et al.(2008) compared three methodologies (multiple discriminant analysis (MDA), logistic regression and hazard models) in order to identify financially distressed companies. The financial ratios are applied to distinguish between Islamic and conventional banks in the GCC region (Olson and Zoubi 2008).

From the above arguments the robust estimator which was entered into LDA to obtain robust linear discriminant analysis (RLDA) where employed to improve the LDA accuracy on classification, investigating the performance predicting bankruptcy and companies distress ...etc. Also the financial ratios have long been used in various study areas in accounting and finance in multivariate methodologies. All of the previous studies focused on obtaining the significant ratios to distinguish and evaluate the company’s performance. This paper presents the linear discriminant analysis with financial ratios, in particular because almost all of the financial ratios are spread in interval [0, 1], the simulation data is generated where the mean and covariance matrices values are in the interval [0, 1]. Also in the real data the financial ratios is collected from the annual reports of Islamic and conventional banks in Malaysia. Some
values of ratios are located out of interval $[0, 1]$ and these values considered as outliers. Finally the study finding and conclusion discussed in section 5, 6 and 7 respectively.

2 Financial ratios

The financial ratios are the most important tools to investigate features of companies. The data (financial ratios) are collected from balance sheet of Islamic and conventional banks in Malaysia. This section shows how the data of both banks are spread in the interval $[0,1]$ and it is shows soe observations lies outside of the interval, where this observations represents the outliers observations. From the figures (1-2) the first and second part reflects the Islamic and conventional banks data respectively. The columns represent the financial ratios of the banks speared and how the observations concentrate in the interval $[0,1]$.

![Figure 1: spread the data (financial ratios) in the interval [0,1] with outliers.](image)

3 Methodology

The linear discriminant analysis (LDA) dates as far back as Fisher (1936) and is a widely used technique for describe the separation between the groups of observations. In addition it is used to classify a new observation into one of the
known groups. Even though the linear discriminant analysis has more sophistication non-linear classification method, it is still often used and performs well in many applications. Moreover linear discriminant analysis is a linear combination of the measured variables and easy to interpret.

Assume that we have $g$ populations in a $p$-dimensional space, being distributed with centres $\mu_1, \mu_2, \ldots, \mu_g$ and the covariance matrices $\Sigma_1, \Sigma_2, \ldots, \Sigma_g$. Where the prior probabilities of populations $g$ is denoted by $p_j$, for $j = 1, 2, \ldots, g$ with $\sum_{j=1}^{g} p_j = 1$ the most important features of linear discriminant analysis is the common covariance matrix, so the within groups covariance matrix $\Sigma$ is given by pooling the covariance matrix of different groups as fellows.

$$\sum_{pooled} = \frac{\sum_{j=1}^{g} n_j \Sigma_j}{\sum_{j=1}^{g} n_j}$$ \hspace{1cm} (1)

The means and covariance matrices of each group need to be estimated by using the simple or classical means and covariance matrices as shown equation 2. Since the linear discriminant analysis is very sensitive to outliers the groups may have an unduly influence on the classical linear discriminant analysis therefore the classical method is not robust.
Allocate \( x \) to \( \pi_j \) if \( \hat{d}_{k}^{CL}(x) > \hat{d}_{j}^{CL}(x) \) for \( j = 1, 2, \ldots, g, j \neq k \) with

\[
\hat{d}_{j}^{CL}(x) = \bar{x}_{j}^{t} s^{-1} x - \frac{1}{2} \bar{x}_{j}^{t} s^{-1} \bar{x}_{j} + \ln(p_{j}^{CL})
\]

(2)

where \( s \) is the common covariance matrix and \( \bar{x}_{j} \) is the mean of groups, \( p_{j}^{CL} \) is the membership probability \( \pi_{j} \).

The classical linear discriminant analysis is not robust in contaminated sets, it is very sensitive to outliers and it does not perform well with large datasets that led to use robust estimator to obtained means and covariance matrices instead and plugging into equation (2) and yielding a robust linear discriminant analysis rule (RLDAR), presented as.

Allocate \( x \) to \( \pi_j \) if \( \hat{d}_{k}^{RL}(x) > \hat{d}_{j}^{RL}(x) \) for \( j = 1, 2, \ldots, g, j \neq k \) with

\[
\hat{d}_{j}^{RL}(x) = \bar{\mu}_{j}^{t} \Sigma^{-1} x - \frac{1}{2} \bar{\mu}_{j}^{t} \Sigma^{-1} \bar{\mu}_{j} + \ln(p_{j}^{RL})
\]

(3)

where \( \Sigma \) is the common covariance matrix and \( \mu_{j} \) is the mean of groups, \( p_{j}^{RL} \) is the membership probability \( \pi_{j} \). For both classical and robust methods \( p_{j} \) are estimated the two popular method. Either the \( p_{j} \) are considered to be constant for all populations, yielding \( p_{j} = \frac{1}{n_{j}} \) for each \( j \), otherwise they estimated relatively proportion of the number of observations in each group to the total number of observations in all groups yields \( p_{j} = \frac{n_{j}}{n} \) for each \( j \). Where \( n \) is the total number of observations in all groups \( \pi_{j} \),

For robust linear discriminant analysis we are going to use the Minimum Covariance Determinant estimator (MCD) Rousseeuw and Driessen (1999), so based on the initial estimator \( \mu_{j,0} \) and \( s_{j,0} \), the robust distance for each observations \( x_{ij} \) of group \( j \) is computed Rousseeuw and Zomeren (1990).

\[
RD_{ij}^{0} = \sqrt{(x_{ij} - \hat{\mu}_{j,0})^{t} s_{j,0}^{-1} (x_{ij} - \hat{\mu}_{j,0})}
\]

(4)

For each observation in group \( j \) let.

\[
\omega_{ij} = \begin{cases} 
1 & \text{if } RD_{ij}^{0} \leq \sqrt{\chi_{p,0.975}^{2}} \\
0 & \text{Otherwise}
\end{cases}
\]

(5)
The MCD estimator for group $j$ is obtained as the means $\hat{\mu}_{j,MCD}$ and the covariance matrix $\hat{\Sigma}_{j,MCD}$, as shown by Croux and Haesbroeck (1999). The means and covariance matrix allows flagging the outliers in the data set and obtaining more robust estimates of the membership probabilities as follow. First compute the robust distance for each observation $x_{ij}$ from group $\pi_j$.

$$RD_{ij} = \sqrt{(x_{ij} - \hat{\mu}_{j,MCD})^t \Sigma_{j,MCD}^{-1} (x_{ij} - \hat{\mu}_{j,MCD})} \quad (6)$$

Here $x_{ij}$ is an outlier if and only if $RD_{ij} > \sqrt{\chi^2_{p,0.975}}$

Let $\hat{n}_j$ denote the number of non-outliers in group $j$ and $\hat{n}_j = \sum_{j=1}^{g} \hat{n}_j$, then the membership probabilities of group $j$.

$$\hat{p}_j^R = \frac{\hat{n}_j}{\hat{n}} \quad (7)$$

in studying the performance of linear discriminant analysis with financial ratios in interval $[0,1]$ and as mentioned previously the linear discriminant analysis is highly influence to outliers, when the real data might have outliers, it comes during the collecting of the data or classification into groups, so by estimating the misclassification probabilities of all groups with weights equal to membership probability we can study the performance of linear discriminant analysis in simulation study and real example as well, where the misclassification probabilities presented as.

$$MP = \sum_{j=1}^{g} \hat{p}_j^R MP_j \quad (8)$$

The fast way to study the performance of three techniques is statistical programming, so we used S-PLUS software because it includes the FAST-MCD algorithm. The code of LDA and misclassification probability is written in S-PLUS.
4 Simulation study

In the case of linear discriminant analysis there are three parameters $\mu$, $\Sigma$ and $P_j$. Which is supposed that all populations have the means $\mu_j$ and the common covariance matrices $\Sigma$ which need to estimate from the data. Robust linear discriminant analysis (RLDA) in formula (3) is used to estimate these parameters. The RLDA is applied by utilizing the Minimum Covariance Determinant (MCD) to obtain the parameters. We adopt three methods that was used by "Hubert" to estimate the means and common covariance matrices for all groups with raw and reweighted versions. The first approach is straightforward and it has applied by Chork and Rousseeuw (1992) $\mu_j$ and $\Sigma_j$ are obtained by pooling the covariance matrices $\Sigma_j$ as:

$$\hat{\Sigma}_{PCOV} = \frac{\sum_{j=1}^{l} n_j \hat{\Sigma}_{j,MCD}}{\sum_{j=1}^{l} n_j}$$ (9)

When the RLDA based on $\mu_{j,MCD}$ and the pooled covariance matrix $\Sigma_{PCOV}$ this approach will denote by pooled covariance PCOV, and PCOV-W for reweighted version.

For the second approach the idea is based on pooled the observations instead of the groups covariance matrices. This approach is one of the proposals of He and Fung (2004) who use S-estimator and it was adapted by Hubert and Driessen (2004), to simplify the approach the notation of three groups same number of groups in simulation study and for farther groups are in the same pattern. Suppose there are three samples $A = (a_{11}, a_{21}, \ldots, a_{n_{1,1}})$, $B = (b_{12}, b_{22}, \ldots, b_{n_{2,2}})$, $C = (c_{13}, c_{23}, \ldots, c_{n_{3,3}})$. Let $\mu_A, \mu_B$ and $\mu_C$ are the locations estimator of the populations as reweighted MCD, the pooled and shift the observations as follows:

$$Z = (z_1, z_2, \ldots, z_n) = (a_{11} - \mu_A, a_{21} - \mu_A, \ldots, a_{n_{1,1}} - \mu_A, b_{12} - \mu_B, b_{22} - \mu_B, \ldots, b_{n_{2,2}} - \mu_B, c_{13} - \mu_C, c_{23} - \mu_C, \ldots, c_{n_{3,3}} - \mu_C)$$

The observations at this approach are pooled instead of pooling the covariance matrices, so the RLDA is denoted by POBS and POBS-W for reweighted version.
The third approach is mixed between two previous methods. It’s based on the MCD estimator and aimed at finding a fast approximation to the Minimum Within-group Covariance Determinant (MWCD) criterion of (Hawkins and Malachian 1997). Instead of doing same adjustment for each group, they proposed to find \( h \) out of all observations set of size \( n \) where covariance matrix \( \Sigma_H \) of \( h \) has minimal determinant, then covariance matrix of \( H \) ( \( h \) out of \( n \) ) is obtained. The approaches for the three groups are described below.

1. Estimate the centres of groups
2. Shift and pool the observations to obtain \( z \)’s same as in the second approach then by using the MCD estimator.
3. Let \( H \) be the \( h \)-out of the \( n \) based on minimizes MCD estimator.
4. Partition the subset \( H \) to \( H_A, H_B \) and \( H_C \) that contain the observations from \( A, B \) and \( C \) respectively.
5. Estimate the mean of all groups as \( \mu_A, \mu_B \) and \( \mu_C \) as well.

The RLDA at this approach named based on the Minimum Within-group Covariance Determinant as MWCD and WMCD-W for the reweighted version.

5 Simulation result

During the simulation study we compared the performance of three methods of RLDA with raw and reweighted versions for each of the methods. The methods were compared using sitting as:

\[
A \Rightarrow \pi_1 : 100N_3(\mu_{A1}, \delta) \\
\pi_2 : 100N_3(\mu_{A2}, \delta)
\]

\[
B \Rightarrow \pi_1 : 100N_3(\mu_{B1}, \delta) + 20N_3(5, (1/0.25^2)\delta) \\
\pi_2 : 50N_3(\mu_{B2}, \delta) + 10N_3(-4, (1/0.25^2)\delta)
\]

\[
C \Rightarrow \pi_1 : 180N_3(\mu_{C1}, \delta) + 36N_3(0.25\mu_{C1}, (1/0.25^2)\delta) \\
\pi_2 : 60N_3(\mu_{C2}, \delta) + 12N_3(-1/0.25)\mu_{C1}, (1/0.25^2)\delta)
\]

\[
D \Rightarrow \pi_1 : 150N_3(\mu_{D1}, \delta) + 15N_3(1/0.075)\mu_{D1}, (1/0.075)\delta) \\
\pi_2 : 60N_3(\mu_{D2}, \delta) + 5N_3(-1/0.075)\mu_{D1}, (1/0.075)\delta)
\]

\[
E \Rightarrow \pi_1 : 160N_3(\mu_{E1}, \delta) + 30N_3(5\mu_{E1}, 0.25^2\delta) \\
\pi_2 : 80N_3(\mu_{E2}, \delta) + 15N_3(-10\mu_{E2}, 0.25^2\delta)
\]
Table 1: Table 1:- Mean $\mu$ and Standard deviation (SD) for the Misclassification Probability estimates for the raw and reweighted robust linear discriminant analysis based on 500 replications.

<table>
<thead>
<tr>
<th>Case</th>
<th>Raw COV</th>
<th>Raw OBS</th>
<th>MWCD - COV</th>
<th>MWCD - OBS</th>
<th>Raw - W COV</th>
<th>Raw - W OBS</th>
<th>MWCD - W COV</th>
<th>MWCD - W OBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.454</td>
<td>0.446</td>
<td>0.249</td>
<td>0.451</td>
<td>0.452</td>
<td>0.229</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.035</td>
<td>0.032</td>
<td>0.081</td>
<td>0.032</td>
<td>0.032</td>
<td>0.087</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.290</td>
<td>0.285</td>
<td>0.388</td>
<td>0.280</td>
<td>0.281</td>
<td>0.355</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.044</td>
<td>0.039</td>
<td>0.067</td>
<td>0.036</td>
<td>0.036</td>
<td>0.013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.370</td>
<td>0.359</td>
<td>0.299</td>
<td>0.330</td>
<td>0.330</td>
<td>0.284</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.094</td>
<td>0.087</td>
<td>0.062</td>
<td>0.080</td>
<td>0.079</td>
<td>0.015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.341</td>
<td>0.329</td>
<td>0.274</td>
<td>0.291</td>
<td>0.289</td>
<td>0.260</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.052</td>
<td>0.037</td>
<td>0.002</td>
<td>0.029</td>
<td>0.027</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.332</td>
<td>0.329</td>
<td>0.400</td>
<td>0.317</td>
<td>0.317</td>
<td>0.374</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.031</td>
<td>0.029</td>
<td>0.003</td>
<td>0.026</td>
<td>0.026</td>
<td>0.018</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We assumed equal membership probabilities in all cases. Moreover the $\delta$ is a three dimensional covariance matrix and $\mu_j$ is the center of all groups in all cases. As mentioned previously the values of the means and covariance matrix are in the interval $[0,1]$. The data is generated in different cases. Case $A$ is generated from uncontaminated populations, but in cases $B, C, D$ and $E$ are generated from contaminated populations also, they are generated with different percentage of outliers'. The experiment is replicated 500 times. Table (1) shows the means and standard deviations of misclassification probabilities of the raw and reweighted discriminant rules.

From table (1) we can see the reweighted versions clearly better than the raw versions because of increasing the efficiency of the linear discriminant analysis rules. Also we can conclude that the MWCD method performance is the most accurate in difference between raw and reweighted versions obtained compared to PCOV and POBS methods. For PCOV and POBS methods there are comparable performances but at MWCD the performance is not comparable with both PCOV and POBS. Case $A$ of MWCD at reweighted version has the best performance comparison with other cases. For case $B$ and $E$ have almost same results in same version, $C$ and $D$ have almost same.

Although from the difference of the size of groups and the interval of values of the means and covariance matrices, also the percentage of the outliers at the sitting of this simulation study, but the misclassification probability of the reweighted versions performance for all linear discriminant analysis better than raw versions. This result is in agreement with simulations study of Hubert and Van (2004).
As mentioned before the outcome of simulation study is obtained through five cases and each case replicated 500 times for three methods at raw and reweighted versions. Figure (3) represent the frequency of the values of the result of misclassification probabilities for 500 replications at case (A). The first three graphs are taken for three methods of the raw version and the other three are taken for three methods from the reweighted version.
The first part of figure (2) is representing the values of 500 replications under PCOV method. We can see that most values are between 0.4 and 0.5 and concentrated at 0.45 that is mean same result at table (2) at PCOV column. POBS, PCOV-W, and POBS-W graphs the misclassification probability values are concentrated around 0.45 same values as in table (2). For the MWCD and MWCD-W methods however the values are concentrated around 0.23 same results from the table (2).

6 Real Data

The data was collected from the annual report of the Islamic and Conventional banks in Malaysia. We obtained the financial ratio regardless for which part they belong to. Three financial ratios were collected as variables in data set, also we collected three variables is equal to the number of variable in the simulation study. Where the variables are.

<table>
<thead>
<tr>
<th>Table 2: Table 2:- Definitions of three financial ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Total Assets to Net Worth</td>
</tr>
<tr>
<td>2. Sales to Fixed Assets</td>
</tr>
<tr>
<td>3. Asset Turnover</td>
</tr>
</tbody>
</table>

The total data set contain 178 observations, 70 observations collected from Islamic banks and 108 observations collected from conventional banks both of them under Central bank of Malaysia. The observations have been classified in two groups Islamic and conventional data group and both of them collected in different period based on the bank’s date of processing (e specially Islamic banks), the maximum period is five years from 2005 until 2009, where the number of years data are different between the banks.

In the example we used the same method that used in simulation study (PCOV, POBS, MWCD) for same versions raw and reweighted to compare the performance of all those methods through the real data and compare it with the performance of the simulation study.

From table (3) we can conclude that the performance of reweighted versions is better than the raw versions for all methods (PCOV, POBS, MWCD). Also the most accuracy is obtained from MWCD at both raw and reweighted
Table 3: The Misclassifications Probability estimates for raw and reweighted robust linear discriminant analysis rules based on the real data (the financial ratios for Islamic and Conventional banks in Malaysia)

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>Reweighted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>0.2051</td>
<td>0.2577</td>
<td>0.0520</td>
</tr>
<tr>
<td></td>
<td>0.1582</td>
<td>0.2179</td>
<td>0.0482</td>
</tr>
</tbody>
</table>

versions. All three methods in raw and reweighted versions do not have a good comparable performance. MWCD method gave the best result comparing with other methods in raw and reweighted versions. Similarly at the level of raw and reweighted versions the MWCD-W is gave the best result. For real data example and the simulation study in previous section the raw and reweighted versions gave the same result, where reweighted versions is performed better than raw versions. For all methods (PCOV, POBS, MWCD) the result were same where MWCD for reweighed version result is better than PCOV and POBS methods. In other case PCOV and POBS gave relatively same result different from MWCD different level.

7 Conclusion

In this paper we studied the performance of robust linear discriminant analysis by using the robust MCD-estimator with three methods to obtain the initial estimate for groups mean and common covariance matrix. The membership probability is estimated by robust method and the misclassification probabilities as well. Through the simulation study the data is generated by using mean and covariance matrix values between \([0, 1]\) interval, because the values of the financial ratios data are in the interval. The outcome of the simulation study clearly showed how the reweighted versions is better than the raw versions for all (PCOV, POBS, MWCD) methods in all cases (A, B, C, D, E) as well, and the MWCD approach is better than the other approaches in investigate the misclassification probabilities. Finally to investigate the outcome in the simulation study we applied the robust linear discriminant analysis rules to real data sets (financial ratios of Islamic and Conventional banks of Malaysia), where the outcome of our study in real data showed same result in simulation study.
8 Acknowledgement

The authors would like to thank school of mathematics at Universti Sains Malaysia "USM" and the USM short term grant project number "PMATHS.6130090" for providing financial support for this study.

References


Received: June, 2012