Using the High Robustness Discriminant Analysis in Classification and Predictions (A Comparative Study)

Khalid Hyal Hussain
Dept. of Statistics, University of AL- Qadisiya, khaledhyal@outlook.com

Hassan S. Uraibi
Dept. of Statistics, University of AL- Qadisiya, hassan.uraib@qu.edu.iq

Abstract

In this study, one of the important multivariate statistical analysis methods called (discriminant analysis) was reviewed, which is used for classification and prediction through the linear discriminant function of (Fisher). However, the main problem that the study addressed is how to use the discriminant analysis when one of the basic hypotheses is violated due to the presence of outliers. The main objective was how to classify and predict when there are outliers in the data under study. Therefore, a robust and resistant method with a high breaking point was proposed to obtain accurate results in classification and prediction. The robust method was compared with the classical method, and the study concluded that the robust method has a high ability to analyze data in the presence of outliers values, as the immune function gave results with high efficiency and the classification error was very low.

Key words: *Discriminant analysis, classification, Robustness.*

1. INTRODUCTION

Statistics is a set of methods and methods that deal with the data of various phenomena in terms of collecting, tabulating, analyzing and extracting the results of the analysis to make a specific decision according to the type of study. Data analysis processes are conducted using special statistical methods called (Statistical Analysis Methods). They are classified into three types:

1- Univariate models: It is the model that studies the existence of only one independent variable in the data and its effect on the dependent variable must be known, and there are no other effects or relationships within the data, such as the following model:

$$y = \alpha + \beta x + \varepsilon \qquad (1)$$

Where:

y: is the dependent variable

x: the independent variable

α: fixed limit

β:is a parameter of the distribution

ε: the random error term of the model

2- Bivariate models: These models are used when there are two explanatory variables to measure their effect on the dependent variable as:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (2)$$

3-Multivariate models: It is considered one of the most complex models because it is used when there are many independent variables and to measure their effect on the dependent variable as:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (3)$$

Where (p) is the number of independent variables, i=1…,n

Where n the sample size.

2. Multivariate analysis:

Statistical methods are classified according to the number of variables included in the model. Multiple models study a set of variables of a phenomenon or research that follow the same distribution. Therefore, multivariate statistical analysis is a set of methods and methods of statistical analysis that use several variables at the same time. But it includes all types of data and the type of study or phenomenon involved accordingly. The term multivariate in the statistical literature refers to all statistical methods that deal with a mathematical function containing more than two variables, and these variables follow a multivariate normal distribution and share many properties.

2.1 conditions of multivariate model:

There are many conditions and assumptions that characterize the methods of multiple statistical analysis, such as:

1- The linear relationship between the variables (linearity)

2- Variation homogeneity of the samples.

3- No problem of multicollinearity.

4-No measurement and aggregation errors.

6- No outliers observation.

7- The random error is normally distributed with zero mean and constant variance:

$\varepsilon \sim N(0,\sigma^2)$

8- Errors are independent in each experiment

9- No autocorrelation problem

2.2 Kinds of multivariate statistical analysis

There are many methods of multivariate statistical analysis, such as:

a- Principal Components Analysis

b- Factorial Analysis

c- Cluster Analysis

d- Discriminant analysis

2.2.1 Discriminant Analysis

Discriminant analysis is one of the methods of multivariate statistical analysis that is concerned with classification and prediction by distinguishing between groups using the discriminant function. The discriminant function is a linear combination between the independent variables and the independent variable that increases the variance between the groups.

2.2.1.1 Concept of Discriminant Analysis

It is a powerful statistical method for characterizing the elements of society, and the data must be divided into two or more separate groups, to know the boundaries between the data, and to establish a specific rule to know that any new element belongs to the appropriate group.

2.2.1.2 Assumptions of Discriminant analysis

Some assumptions and conditions that must be applied when using discriminant analysis, such as:

1- The independent variables must follow the multiple normal distribution, and this condition is met when the sample size is increased, as in the theory of central purpose according to (Buyukozturk 2008).

2- The groups should not be overlapping, and their number should be greater than 2, and their sizes should be consistent.

3-Stratified sampling was used to make the selection of observations independent. (Steven 1996)

4-The explanatory variables (X1, X2, …, XP) are independent of each other.

5-The number of independent variables is less than the number of observations (P<n), and if their number is large.

6- there is homogeneity in the variance of groups. groups.

7-The errors (residuals) should be the normality distributed of errors with an expectation (zero) and variance (σ2).

8- the randomly drawn of sample.

9- There are no outliers in the data.

2.2.1.3 The goal of using discriminant analysis

There are several purposes for discriminant analysis:

1-known as the differences between the groups.

2-find the optimal rule for discriminant between samples.

3-know which independent variables are important.

4-shrinking variables with weak correlation.

5- Classifying observations into groups.

6-Predicting of new observations to the appropriate groups.

7-measures the classification accuracy.

2.2.1.4 Application of D.A.

Previous studies used the method of discriminatory analysis in many fields and phenomena, such as the educational aspect, where the method of discriminatory analysis was used to arrange students on a number of disciplines according to their mental abilities and level of intelligence. As for the agricultural and biological aspect, the discriminant analysis method was used to classify some rare agricultural crops and determine the strain, in addition to some small organisms according to the category to which they belong. As for the medical aspect. There are many applications of the discriminatory analysis method in the medical aspect, such as classifying the patient's condition into (simple, moderate, and severe). Perhaps one of the most important applications of discriminatory analysis is predicting financial failure and classifying institutions or banks into non-performing and non-performing ones using ratios and financial indicators.

2.2.1.5 The linear Discriminant Function

There are many kinds of functions of discriminant as; linear, logistic and quadratic discriminant analysis function. each function has some assumptions for use, depending on the type of study and data so We will introduce some popular functions of discriminant analysis. (R. Fisher 1936) suggested the Linear Discriminant Function (LDF), Where he assumed that this function is the best way to classify and discriminant in scientific issues, which is as in the following formula:

$$Z = x'\Sigma^{-1}(\bar{x}_1 - \bar{x}_2) \qquad (4)$$

Where:

$\bar{x}_1, \bar{x}_2$ is the mean of the first and the second group respectively.

Σ is the (Var-Cov) matrix.

The cutter point is:

$$L = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' s^{-1}(\bar{x}_1 - \bar{x}_2) \qquad (5)$$

Also, we can compute the cutter point from:

$$L = \frac{\bar{y}_1 + \bar{y}_2}{2} \qquad (6)$$

Where:

$$\bar{y}_1 = \bar{x}_1' s^{-1}(\bar{x}_1 - \bar{x}_2) \qquad (7)$$

$$\bar{y}_2 = \bar{x}_2' s^{-1}(\bar{x}_1 - \bar{x}_2) \quad (8)$$

The classification decision will be as follows:

Classified to the first group when $L \leq Z_{x_0}$
Classified to the second group when $L > Z_{x_0}$

2.2.1.6 Misclassification errors

Classification error is an important for measuring the efficiency of the discriminant function, where the purpose of is to obtain the lowest possible classification error, which is the probability of classifying a particular observation into the first group, but in fact it belongs to the second group, and vice versa, assuming that the data follow the normal distribution or approach it in the case of large samples according to the central goal theory. The classification error is calculated based on the standard normal distribution function. In the case of two groups, the probability of misclassifying any observation into the first group, but belonging to the second group, is:

$P_{1\,2}$ = P {classifying x to be from group (1) / x is from group (2)}

$$= \emptyset\left(-\frac{D^2}{2}\right) \quad (9)$$

This value is calculated from the standard normal distribution tables, and D^2 is the Mahalanobis statistic which is:

$$D^2 = (\bar{x}_1 - \bar{x}_2)' \Sigma^{-1}(\bar{x}_1 - \bar{x}_2) \quad (10)$$

Where:

$\bar{x}_1, \bar{x}_2$:is the means of the first and the second groups respectively

Σ: is the (Var-Cov) matrix.

2.2.1.7 The robust discriminant analysis

Linear discriminant analysis assumes that the variances are homogeneous and that the relationship is linear between the variables, as well as the absence of abnormal values because it leads to a violation of the hypothesis of the normal distribution according to (Hamble 1974)

Therefore, it was necessary to fortify the traditional discriminatory analysis method to obtain a discriminatory function that is resistant to the presence of outliers and gives accurate results in classification and prediction. A robust method was proposed to strengthen the discriminant analysis method to resist outliers. In order to fortify the linear discriminant analysis method from the influence of outliers, Jana has to adopt a robust method to reduce the effect of those outliers in the data under study. It was necessary to resort to converting the data using a specific location and measurement matrix (RMVN), and then employing it to fortify the Mahalanobis Distance method according to (Uraibi 2017), as well as to take advantage of the (FCH) estimates, which represent the fastest estimates of High breaking point (Olive & Hawkins 2010). The new matrix is used to weight the estimators of the multiple normal distribution using an efficient algorithm that has a high breakdown point.

3. Application side

3.1 the sample of study

The practical side of this study was adopted by taking a random sample of (13) banks from the Iraq Stock Exchange, the most heavily traded for the period (2010-2017), where the total number was (48) banks, and the banks included in the study sample are (Al-Ahly, Assyria, Babylon, Baghdad, Elaf, Al-Itman, Investment, Al-Iraqi, Al-Khaleej, Al-Khaleej - Mansour, Sumer, Al-Tijari, The National Bank). (28) financial ratios were collected in this thesis, which are the independent variables of the study. Therefore, the general discriminatory model of the study will be in the following form:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{28} X_{28} + \varepsilon \qquad (11)$$

Where ($\beta_1, \beta_2, \ldots, \beta_P$) are called discriminatory coefficients.

(p) is the number of predictive variables which is equal to (28) in this study.

(ε) represents the classification error of the discriminatory model.

3.2 Separate the data into two groups

The discriminatory analysis method depends on dividing the study sample into two groups or more. Therefore, a new method was proposed to classify the data into two groups, which is to choose one of the banks included in the study and consider it a standard bank for the rest of the banks. We can identify the defaulting and non-performing banks, compared to the performance of this bank. The process of selecting such a bank cannot be subject to personal or speculative criteria, and therefore we resorted to selecting the standard bank through the following algorithm:

1- take one of the study banks as the standard bank.

2- measures the standard average for each financial ratio of the standard bank and for the years under study (8 years).

3- Shown the financial ratios of other banks to the years with the standard averages of those ratios in step (2). If the bank's financial ratio is greater than the standard average, it shall be replaced by the number (1), otherwise it shall be replaced by the number (0).

4- The sum must be obtained for each financial ratio (as a binary variable) for eight years. If it is greater than or equal to (4), the sum is replaced by the value (1), and in others the number (0) is written. Then we get (28) values, and by taking the sum of the values, a final decision can be obtained to describe the concerned bank as either defaulting or not defaulting. If the sum is greater than or equal to (14), the bank is considered non-performing, otherwise it is considered non-performing.

5- Repeat the previous steps and each time we take out a bank.

6- the standard bank is from the results, provided that this bank achieves a balance between troubled and non-performing banks, as we will see later.

The results presented in Table (1) include determining the number of non-performing banks by choosing an assumed standard bank based on the mean of the financial indicators.

Table (1) choosing the standard banks

The assumed bank	Num. failed banks	Num. Successful banks
AL-Ahly	3	9
Ashur	6	6
Babel	0	12
Baghdad	8	5
Eilaf	0	12
AL-Eitiman	10	2
AL-Estithmar	10	2
AL-Iraqi	5	7
AL-Khaleeg	9	3
AL-Mansour	3	9
Sumer	7	5
AL-Tijary	12	0
AL-Watany	9	3

From the results above, we note that when choosing (Al-Ahly Bank) as a criterion, the number of defaulted banks is (3), while the number of defaulted banks is (9), and this is unbalanced in the results, as the two groups

differ greatly, so they were excluded. Whether in the case of choosing a bank (Babylon) or (Elaf) as a standard bank, we will not get a group of troubled banks. We may not get non-performing banks when choosing the commercial bank that shows the best performance among the banks of the study sample. However, we note that the Ashur Bank presents a balanced situation, because it divides the banks into two groups equally, unlike the rest of the banks, which diversified into classifying banks into two groups.

Therefore, the results of using the arithmetic average of the financial indicators of the standard bank should be viewed to find out the defaulting banks from the defaulting banks, as in Table No. (2).

Table (2) separate the banks into two groups

failed banks	Successful banks
Babel	AL-Ahly
Baghdad	AL-Watany
Eilaf	AL-Tijary
AL-Iraqi	AL-Eitiman
AL-Mansour	AL-Estithmar
Sumer	AL-Khaleeg

3.3 The results of the classical discriminant analysis model (LDA)

The results of the linear discriminant function (LDA) analysis were obtained using the programming language (R. Program) and the software package (klaR). The sample of troubled and non-performing banks in Table (3-6) is considered to belong to two different communities. These results were applied as inputs to the discriminant analysis function (LDA) in the mentioned package and according to the financial indicators for each bank. (8) samples were entered for each of the (13) banks. The total sample size was (104) observations from the total banks of the study sample. The default cases for each bank were classified according to the years according to the aforementioned algorithm. The results showed that there were (48) failure cases with a probability of (0.46) and (56) non-default cases with a probability of (0.54). The results of the conventional discriminant analysis can be summarized as in this step.

3.3.1 The important variables

To obtain a strong discriminating equation, the important financial indicators that affect the discrimination process must be extracted. This method is implemented using forward progressive regression and using Wilks' Lambda test confined between ($1 \geqslant \Lambda \geqslant 0$), where the closer we get to (1), the more influential the variable is in the model. The test (F. statistics and P. value criterion) was used at the level of significance (0.05) to choose the significant percentages. As in Table (3).

Table (3) presents the results of forward gradual regression using Wilks' Lambda statistic, Fisher test and P.value.

Selective variables	Wilks. Lambda	P.value	F. value
X_{24}	0.864815	0.000123	**15.94425**
X_{19}	0.793577	8.5E-06	**13.13591**

X_{28}	0.758398	4.02E-06	**10.61895**
X_{17}	0.726403	1.95E-06	**9.321973**
X_{21}	0.705957	1.79E-06	**8.163738**

We note the results presented in Table (3). There are (5) variables that are among the important ratios in the discriminatory equation, where the highest value of (Wilks' Lambda) statistic is for the variable (X24) was (0.864815), and thus it is one of the most influential variables in the discriminatory model, and then the rest of the variables come after it in terms of importance. Thus, the discriminatory model becomes according to the variables included in it in terms of importance according to the following formula:

$$z = \beta_{17}X_{17} + \beta_{19}X_{19} + \beta_{21}X_{21} + \beta_{24}X_{24} + \beta_{28}X_{28}$$

3.3.2 Discriminant coefficients

After identifying the important financial indicators that affect the classification processes of the (LDA) method, we find the averages of those indicators for stumbling and non-stumbling banks, as well as the discriminatory coefficients of the discriminatory model., as shown in Table (4):

Table (4) Means of the significant variables stumbling and non- stumbling banks and the discriminatory coefficients of the results of (D.A)

discriminant coefficients	Average indicators for successful banks	Average indicators for failed banks	important variables
13.61	0.96	0.92	X_{24}
6.25	0.095	0.08	X_{19}
0.01	41.5	11.93	X_{28}
0.37	0.43	0.519	X_{17}
-2.04	0.46	0.49	X_{21}

After estimating the discriminatory coefficients as above, we wrote a discriminatory function to predicting the classification of the new observations as follows:

$$y = 0.37x_{17} + 6.25x_{19} - 2.04x_{21} + 13.61x_{24} + 0.01x_{28}$$

We note that the variable (x_{24}) has the largest discriminant coefficient (13.61), which means that it is the most influential variable in the model as it is directly proportional. The variable (x_{21}) has a negative value (-2.04), which means that its effect on the predictive value is inversely proportional.

3.3.3 The cut of point

The cut-off point is the most important predictive step in classifying new views into one of the two groups (distressed and non-distressed), which is calculated according to the following formula:

$$L = \frac{\bar{y}_1 + \bar{y}_2}{2} = \frac{0.417 + (-0.668)}{2} = -0.1255$$

Predictive values were calculated for all banks according to years, as shown in Table (3-9):

Table (5) The predictive value of the discriminatory function in the discriminatory equation and the classification of banks into defaulting and non-performing banks by years

Classify.	Disc.val.	year	Bank	classification	Discriminant value	year	Bank	Clas.	Disc.val.	year	Bank	Classify.	Disc.val.	year	Bank
S	2.577	٢٠١٦	Estithmar	S	3.446	٢٠١٤		S	-0.414	٢٠١٢		S	-0.019	٢٠١٠	
S	-0.549	٢٠١٧		F	-0.587	٢٠١٥		S	-0.161	٢٠١٣		S	-0.548	٢٠١١	
S	-0.252	٢٠١٠		F	-0.945	٢٠١٦		S	-0.158	٢٠١٤		S	0.008	٢٠١٢	
S	-0.506	٢٠١١		F	-0.939	٢٠١٧	Ahly	F	-1.399	٢٠١٥	Iraqi	S	0.181	٢٠١٣	Babel
S	-0.383	٢٠١٢		S	0.301	٢٠١٠		F	-0.687	٢٠١٦		S	-0.023	٢٠١٤	
S	-0.227	٢٠١٣		S	0.146	٢٠١١		F	-1.136	٢٠١٧		S	-0.051	٢٠١٥	
S	-0.398	٢٠١٤	Khaleeg	S	0.163	٢٠١٢		S	0.21	٢٠١٠		F	-1.365	٢٠١٦	
S	-0.158	٢٠١٥		S	0.041	٢٠١٣	Ashur	S	0.109	٢٠١١		S	-0.303	٢٠١٧	
S	-0.457	٢٠١٦		S	0.041	٢٠١٤		S	-0.092	٢٠١٢	Mansour	F	-0.682	٢٠١٠	Baghdad
S	-0.487	٢٠١٧		S	-0.344	٢٠١٥		S	0.415	٢٠١٣		F	-0.927	٢٠١١	
S	1.58	٢٠١٠		F	-1.259	٢٠١٦		S	0.457	٢٠١٤		F	-0.916	٢٠١٢	
S	1.219	٢٠١١		F	-1.397	٢٠١٧		S	0.19	٢٠١٥		F	-1.277	٢٠١٣	
S	1.012	٢٠١٢		S	1.689	٢٠١٠		S	-0.282	٢٠١٦		F	-2.746	٢٠١٤	
S	1.368	٢٠١٣		S	2.137	٢٠١١	Eitiman	S	0.161	٢٠١٧		F	-2.299	٢٠١٥	
S	0.834	٢٠١٤	Tijary	S	0.978	٢٠١٢		S	-0.244	٢٠١٠		F	-1.697	٢٠١٦	
S	0.87	٢٠١٥		S	1.559	٢٠١٣		S	-0.129	٢٠١١		S	-0.386	٢٠١٧	
S	-0.139	٢٠١٦		S	2.069	٢٠١٤		S	0.176	٢٠١٢	Sumer	F	-0.662	٢٠١٠	
S	0.039	٢٠١٧		S	1.947	٢٠١٥		S	1.219	٢٠١٣		S	-0.116	٢٠١١	
S	-0.328	٢٠١٠		S	0.79	٢٠١٦		S	-0.476	٢٠١٤		F	-0.886	٢٠١٢	
S	0.807	٢٠١١		S	1.002	٢٠١٧		S	1.172	٢٠١٥		F	-1.518	٢٠١٣	
S	1.279	٢٠١٢		S	0.051	٢٠١٠		F	-1.121	٢٠١٦		S	-0.531	٢٠١٤	Eilaf
S	1.325	٢٠١٣	Watany	S	-0.061	٢٠١١		F	-0.644	٢٠١٧		F	-0.995	٢٠١٥	
S	2.451	٢٠١٤		S	0.087	٢٠١٢		S	-0.323	٢٠١٠		F	-1.632	٢٠١٦	
S	2.259	٢٠١٥		S	0.174	٢٠١٣	Estithmar	F	-0.778	٢٠١١	Ahly	F	-1.836	٢٠١٧	
S	2.747	٢٠١٦		S	0.004	٢٠١٤		F	-0.67	٢٠١٢		F	-1.081	٢٠١٠	
S	2.191	٢٠١٧		S	-0.031	٢٠١٥		F	-1.218	٢٠١٣		F	-1.607	٢٠١١	Iraqi

Table No. (5) shows cases of success and failure of banks and for the years under study, where there are eight cases for each bank with the number of years of study. Showing the case of the bank in that year for the purpose of giving a detailed picture of the case of failure and success (F, S) for each year. The average discriminatory values for performing banks was (-0.668) and non-performing banks (0.417). Therefore, the break point is the average of these two averages:

$$L = \frac{\bar{y}_1 + \bar{y}_2}{2} = \frac{0.417 + (-0.668)}{2} = -0.1255$$

The discriminant values that are greater than the point value are considered not tripped, otherwise they are tripped, which means that the classification decision of the new observation (x_0) is:

classified to group 1 when $L \leq Z_{x_0}$

classified into group 2 when $L > Z_{x_0}$

Where(Z_{x_0}) is the value of the discriminatory function when replacing the new predictive value to be classified

3.3.4 Classification error

The efficiency of the discriminant function depends on the classification error where the goal is to obtain the lowest possible error, which is the probability that a given observation is classified into the first group, but belongs to the second.

Referring to the previous table, which shows the predictive values and cases of success and failure for each bank for the years, where the state of the bank is shown in that year for the purpose of giving a detailed picture of failure and success (F, S) for each year. We note that Babel Bank, which was initially classified as distressed based on the arithmetic mean, continued its success years (S) until the year (2016), which resulted in a failure (F). Another example is the Bank of Baghdad. The rating result for the years (2010-2016) was faltering, but the year (2017) gave a success case for this bank. Therefore, we have what is called classification error.

Table (6) shows the errors in predicting financial failure for the total number of banks for the study sample

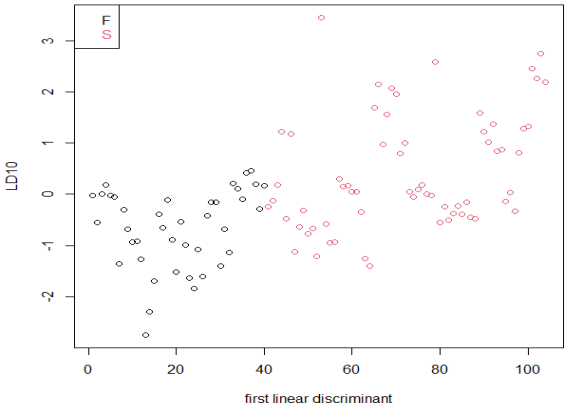
Hypothesis		
Successful	Failure	Prediction
21	19	Failure
54	10	Successful

The classification error rate was (0.30), which is the percentage of cases that did not match the initial classification results. It is an approved ratio and can give acceptable results in classification and prediction, where the classification accuracy rate is (0.70). The failure cases were (19) with non-defaulting cases classified as not defaulting either, which is (54), then dividing the result by the total number of observations, thus the result is:

$$19 + 54 = 73/104=0.7$$

we get the correspondence with the validity of the classification, which is (0.70), then we subtract this ratio from one, we get the classification error of (0.30).

Figure (1-3) predictive values after classification processes



The above figure shows the process of spreading the new observations after classification into two groups, where (S) represents the data that has been classified within the non-performing banks, while (F) indicates the data that has been classified within the non-performing banks. There is a slight overlap between the data at a certain limit. Both groups express a relatively low amount of classification error compared to the correct classification, which makes up the majority of the data.

3.3.5 Testing the ability of the discriminant function

To prove the ability of the discriminant function to distinguish between troubled and non-performing banks, we used Wilks' Lambda test. The basic test hypotheses are to test whether the averages of the two groups of defaulting and non-performing banks are equal or not. The null hypothesis indicates that there are no statistically significant differences between the averages, that is, the inability of the function to discriminate. As for the alternative hypothesis, it indicates that there are statistically significant differences between the means, i.e. the ability of the function to discriminate:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

The results indicate that the value of (Wilks' Lambda) is equal to (0.78) and the value of (P.value) is equal to (0.00015). This means that the null hypothesis is rejected and that the discriminatory function has a high ability to distinguish between groups.

3.4 Results of Robustness Discriminant Analysis (RLDA)

The classical method of discriminatory analysis relied on the use of the arithmetic average of the indicators of the standard bank, which is (Ashur Bank), and then compared them with the corresponding financial indicators and ratios of other banks in order to identify the failing and successful bank. But the arithmetic mean is obviously sensitive in the presence of outliers, as it breaks down in the presence of a single outlier, so we will use a more robust measure based on its accuracy in the case of outliers, which is the median. To reduce the effect of anomalous data, the impervious transformation matrix (RMVN) was used for the location and measurement parameters and was applied to the linear discriminant function, where specific weights were given to the outliers to reduce their impact on the accuracy of the results, then the previous algorithm was applied depending on the financial indicators broker of Ashur Bank, where The median value of each Assyria Bank money ratio was, We rearrange the data so that we give the value 1 for each financial ratio greater or equal to the value of the mediator, and the value 0 for each financial ratio less than the value of the mediator for the Ashur Bank, so the results were as in the following table that represents failures and successes for the National Bank depending on the mediator of the Ashur Bank:

Table (7) Cases of failure and success of the financial ratios of the National Bank in comparison with the mediator of the financial indicators of Ashur bank for the years (2010-2017)

X28	X27	X26	X25	X24	X23	X22	X21	X20	X19	X18	X17	X16	X15	X14	X13	X12	X11	X10	X9	X8	X7	X6	X5	X4	X3	X2	X1	السنة
1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	1	1	0	1	0	0	0	1	0	0	٢٠١٠
1	0	0	1	1	1	0	1	1	1	1	0	0	0	0	0	1	0	1	1	1	1	1	0	0	1	0	0	٢٠١١
1	0	1	0	1	0	0	0	1	0	0	0	0	1	1	1	0	0	1	1	0	1	1	1	0	0	0	1	٢٠١٢
1	0	0	0	1	0	1	0	1	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1	0	1	0	1	٢٠١٣
1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1	0	0	٢٠١٤
1	1	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	٢٠١٥
1	1	1	0	0	0	1	0	1	1	1	0	0	0	1	1	0	0	1	0	0	0	0	1	0	0	0	1	٢٠١٦
1	1	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	٢٠١٧
8	3	2	1	6	1	5	1	8	3	6	0	0	1	3	3	2	0	7	4	3	6	4	4	0	6	0	3	المجموع
1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	1	1	0	1	1	1	0	1	0	0	الحالة
F																												القرار

After applying this mechanism to all banks, two groups of banks were obtained (performing and non-performing) as shown in the table.

Table (8) Classifying the study sample banks into two groups

Failure banks	Succeful banks
Babil	Altijary
Baghdad	Alwatany
Eilaf	Sumer
Iraqi	Eitiman
Almansour	Alestithmar
Alahly	Alkhaleeg

To find out the median value for each financial ratio of the Ashur Bank, on the basis of which the comparison was made as a criterion, it was calculated and placed in Table (9)

Table (9) the median value for each financial ratio of the Ashur Bank

Median	Financial ratios	Median	Financial ratios	Median	Financial ratios
0.629	X_{21}	0.790	X_{11}	0.039	**X_1**
0.962	X_{22}	0.131	X_{12}	0.067	**X_2**
0.534	X_{23}	0.071	X_{13}	0.629	**X_3**
0.933	X_{24}	0.067	X_{14}	8.350	**X_4**
0.629	X_{25}	0.088	X_{15}	2.526	**X_5**
0.043	X_{26}	0.094	X_{16}	1.090	**X_6**
1.400	X_{27}	0.297	X_{17}	0.081	**X_{7-}**
8.199	X_{28}	0.470	X_{18}	0.825	**X_8**
		0.106	X_{19}	0.975	**X_9**
		0.423	X_{20}	0.715	**X_{10}**

3.4.1 Select the important variables

To obtain the best discriminatory equation, the important variables that affect the discrimination process must be extracted. This procedure is done using Stepwise forward selection and using Wilks' Lambda and ((F. Statistics and P.value)) test at a significant level (0.05) to choose the significant percentages, if the (P.value) for that percentage is less than (0.05), then that percentage is considered one of the important percentages that are included in the discriminatory model. The results showed that there are (6) influential proportions in the classification process of the discriminatory hippocampus model, which are as shown in Table (9), which shows the effective variables according to the values of (Wilks. Lambda), the (F. Statistics) test, and the value of (P.value.).

Table (10) the wilks. Lambda test

P.value.	F. Statistics	Wilks. Lambda	Important variables
6.33E-05	17.41712	0.854149	X_{28}
8.22E-06	13.1779	0.793054	X_{20}
5.17E-07	12.51133	0.727093	X_{15}
6.17E-09	13.71833	0.643386	X_{27}
1.04E-10	14.64777	0.5723	X_{6}
5.94E-11	13.31274	0.548405	X_{7}

From the results presented in Table (10), it is clear that there are (6) financial ratios that are considered among the important ratios in the discriminatory equation, where the highest value of the (Wilks' Lambda) statistic for the variable (X_{28}), which represents the index of lift, was (0.854149), and thus it is considered It is one of the most powerful variables in the discriminatory model, followed by the rest of the variables in terms of importance.

Thus, the discriminatory model becomes according to the variables involved in it in terms of importance according to the following formula:

$$Y = \beta_6 X_6 + \beta_7 X_7 + \beta_{15} X_{15} + \beta_{20} X_{20} + \beta_{27} X_{27} + \beta_{28} X_{28}$$

3.4.2 the medians and discrimination coefficients

After selecting the important variables, the medians of each indicator were calculated for troubled and non-performing banks, as well as the discriminatory coefficients of the discriminatory model, as shown in the table.

Table (11) Arguments for each indicator for troubled banks and non-performing banks, in addition to the discriminatory coefficients of the results of the robust discriminant analysis

Discriminant coefficients	Median of non-failure banks	Median of failure banks	Important variables
0.158	7.91	1.996	X_{28}
-1.452	0.244	0.415	X_{20}
133.536	0.017	0.014	X_{15}
-2.944	0.457	0.479	X_{27}
-15.512	0.205	0.217	X_{6}
8.669	0.215	0.224	X_{7}

3.4.3 Cutter point

The cut-off point is the most important predictive step in classifying new observations into one of the two groups (distressed and non-defaulting), which is calculated according to the formula mentioned in equation, which is:

$$L = \frac{\bar{y}_1 + \bar{y}_2}{2}$$

Where $(\bar{y}_1, \bar{y}_2)$ are the average values of the indicators for troubled and non-performing banks. Also, the predictive values of the discriminatory function in the high-immunity

discriminatory model (RLDA) on the basis of which the banks were distinguished into two groups, failure and non-failure.

The average discriminatory values for non-performing banks were (-1.137) and for non-performing banks (0.711). Therefore, the cut-off point is:

$$L = \frac{\bar{y}_1 + \bar{y}_2}{2} = \frac{0.711 + (-1.137)}{2}$$
$$= -0.231$$

Any discriminatory value that passed this point is considered non-failure, otherwise it is failure.

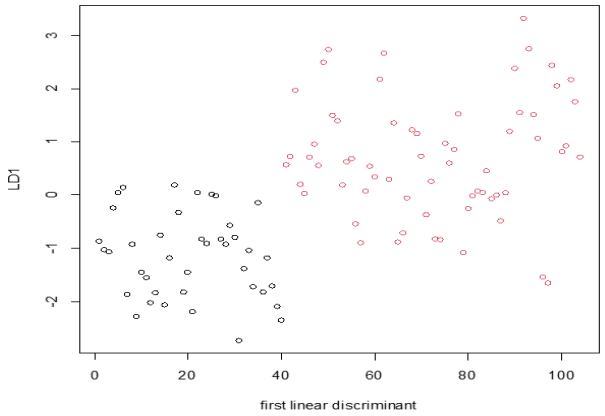
3.4.4 misclassification error

The classification error rate has reached (0.18), which is an approved ratio and can give more accurate results in classification and prediction because it is much less than the error rate of the traditional model. As for the classification accuracy rate, it is (0.82). Actually, which amounts to (31) cases with non-stumbling cases, which were classified as non-stumbling as well, which is (54), then dividing the result by the total number of observations, so the result is 0.82) = 31 + 54 = 85/104) we get the correspondence with the validity of the classification, which is (0.82 Then we subtract this percentage from one, we get the classification error of (0.18).

4. Comparison of methods

We note that there is a significant difference in the classification error rate between the traditional method (LDA), which had a classification error of (0.30), and the robust method (RLDA), which produced a classification error (0.18), which is much less, which gives a clear superiority in accuracy to the latter method. The following figure shows the data intersection processes according to classification accuracy:

Figure (2-3) illustrates the data dissemination process and the accuracy of the classification



The above figure shows the spread of the new data after classification into two groups, where (S) represents the data that was classified within the non-performing banks, while (F) indicates the data that was classified into the non-performing banks. It is clear that there is a slight overlap between the data at a certain limit. The two groups express a classification error, which is much less than the overlap shown in Figure (1-3), which means a significant decrease in the classification error rate in the relatively fortified method compared to the previous method, and this means that the classification accuracy is more accurate in this method.

5. Conclusion

It is clear that the traditional method resulted in a significant classification error, so that its accuracy cannot be relied upon in the results of studies that contain anomalous data. As for the robust method, the superiority is very large, as the classification error is much less than it is in the traditional method. The reason is due to the strong performance of the robust transformation matrix, which handled the existence of outliers based on the (RMVN) algorithm, by giving little weight to outliers

and adopting the median in case of outliers. It is one of the strong measures that are not affected much by outliers and has a high breakdown point. In the traditional method, the arithmetic mean is very sensitive to outliers, as its breakdown point reaches zero, so it may collapse in the presence of one outlier.

Reference

Ahmed, S. W., & Lachenbruch, P. A. (1977). Discriminant analysis when scale contamination is present in the initial sample. In J. Van Ryzin (Ed.), Classification and Clustering (pp. 331-354). New York, NY: Academic Press.

Alrawashdeh, M. J., Muhammad Sabri, S. R., & Ismail, M. T. (2012). Robust linear discriminant analysis with financial ratios in special interval. Applied Mathematical Sciences, 6(121), 6021-6034.

Campbell, N. A. (1982). Robust procedures in multivariate analysis II. Robust canonical variate analysis. Journal of the Royal Statistical Society. Series C (Applied Statistics), 31(1), 1–8. doi: 10.2307/2347068.

Cedric A.B. Smith,1946, Some examples for discrimination.

Chork, C. Y., & Rousseeuw, P. J. (1992). Integrating a high-breakdown option into discriminant analysis in exploration geochemistry. Journal of Geochemical Exploration, 43(3), 191-203. doi: 10.1016/0375-6742(92)90105-H.

Croux, C., & Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. Canad. J. Statist., 29(3), 473–493. doi:10.2307/3316042.

D. J. Hand,1983, A Comparison of Two Methods of Discriminant Analysis Applied to Binary Data, Biometrics, Vol. 39, No. 3 (Sep., 1983), pp. 683-694.

D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema, G. J. Gelpke,1981, Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients, Journal of the Royal Statistical Society. Series A (General), Vol. 144, No. 2 (1981), pp. 145-175.

Edward I. Altman,1968, Financial ratios, discriminant analysis and the prediction of carport bankruptcy, The Journal of Finance, Vol. 23, No. 4 (Sep., 1968), pp. 589-609, America.

Elizabeth A. Chambers and D.R. Cox,1967, Discrimination between alternative binary response models, Imperial college, British.

Feinberg, F. M. (2010). Discriminant analysis for marketing research applications. In N. S. Jagdiesh, & K. M. Naresh (Eds.) Wiley International Encyclopedia of Marketing (Vol. 2). New York, NY: John Wiley & Sons, Ltd doi:10.1002/9781444316568.

I. Akeyede, F.G. Ibi, D. T. Ailobhio,2021, A Discriminant Analysis Procedure for Loan Application using Ranked Data, Asian journal for mathematic sciences, ISSN 2581-3463.

J.A. Anderson,1968, Constrained Discrimination between k Populations, Oxford university, British.

M.L. Tildesley,1921, A first study of the Burmese skull, University College, London.

Maharaj, E. A., & Alonso, A. M. (2014). Discriminant analysis of multivariate time series: application to diagnosis based on ECG signals. Computational Statistics & Data Analysis, 70, 67-87. doi: 10.1016/j.csda.2013.09.006.

Michael Q. Zhang,2000, Discriminant analysis and its application in DNA sequence motif recognition, Briefings in Bioinformatics, Volume 1, Issue 4, 1 November 2000, Pages 331–342.

Mufda J. Alrawashdeh, Taha Radwan and Khalid Abunawas,2018, Performance of linear discriminant analysis using different robust methods, European Journal of pure and applied mathematics, Vol. 11, No. 1, 2018, 284-298.

P.C. Mahalanobis,1936, On the generalized distance in statistics, India.

R.A. Fisher,1936, The use of multiple measurements in taxonomic problem, the Annals of Eugenics, v. 7, p. 179-188 (1936) with permission of Cambridge University, London.

Randles, R. H., Broffitt, J. D., Ramberg, J. S., & Hogg, R. V. (1978a). Generalized linear and quadratic discriminant functions using robust estimates. Journal of the American Statistical Association, 73(363), 564-568. doi: 10.2307/2286601.

Richard J. Beckman and Mark E. Johnson,1981, A Ranking Procedure for Partial Discriminant Analysis, Journal of the American Statistical Association, Vol. 76, No. 375 (Sep., 1981), pp. 671- 675.

Ronald H. Randles, James D. Broffitt, John S. Ramberg & Robert V. Hogg,1978, Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates, Journal of the American Statistical Association, 73:363, 564-568.

Sajobi, T. T., Lix, L. M., Dansu, B. M., Laverty, W., & Li, L. (2012). Robust

descriptive discriminant analysis for repeated measures data.

Computational Statistic and Data Analysis, 56(9).

Shelley B. Bull and Allan Donner,1987, The Efficiency of Multinomial Logistic Regression Compared with Multiple Group, Journal of the American Statistical Association, Vol. 82, No. 400 (Dec., 1987), pp. 1118- 1122.

Wina, Herwindiati, D. E., & Isa, S. M. (2014). Robust discriminant analysis for classification of remote sensing data. In A. Wibisono (Ed.), Proceedings of 2014 International Conference on Advanced Computer Science and Information System (pp. 454-458), IEEE. doi: 10.1109/ICACSIS.2014.7065892.