THE AMERICAN UNIVERSITY IN CAIRO | SCHOOL OF SCIENCES AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND ACTUARIAL SCIENCE

Discriminant and Cluster Analysis

Multivariate Analysis (MACT4233-02)

Dr Ali Hadi

Spring 2023

Laila El Saadawi

900201723

Masa Tantawy

900201312

Table of Contents

Introduction

Counterfeit money, colloquially known as "fake bills," is produced without a country or government's permission. The production of fake bills falls under the crime of fraud or forgery and is punishable by up to twenty years in the U.S. Historically known as "the world's oldest profession", counterfeiting money poses a threat to not only the country's economy and its inflation levels but also to members of its business sector who are not being fairly reimbursed for the goods and services they provide. In the past, identifying counterfeit money was far from an easy task, so bill-makers adopted the habit of creating complex, irreplicable patterns on the gold coins used as currency. As for paper money, imprinting the impression of a unique and intricate figure that was not easily reproduced (such as a leaf on Colonial American money) was common. With the advancement of technology and powerful machinery, counterfeiters have become more skilled in the art of replicating bills that can pass under the radar of most establishments. Yet, with this improvement also came massive advancements in the detection methods to identify even the top-tier counterfeit bills. Simple methods to identify counterfeit bills include assessing texture, printing quality, and lack of detail on the money. Those accustomed to handling money regularly can tell the differences such as the color differences. Subtle variations raise suspicion since all government-printed bills share near identical thickness and dimensions.

In this project, the dataset (fakebills.csv) provides details about the dimensions of various genuine and fake U.S. dollar bills. By applying multivariate analysis techniques, we will be able to single out any confounding outliers as well as test the statistical make-up of the data. Analyzing the imperceivable differences that could distinguish genuine money from its counterfeit counterpart will in turn greatly help improve the accuracy of models that accurately distinguish between the fake and authentic dollar bills.

Data Description

Data Source: Kaggle

https://www.kaggle.com/datasets/alexandrepetit881234/fake-bills

The dataset describes the dimension of different fake and genuine bills.

Number of observations and variables: 1500 observations and 7 variables (Including a boolean variable representing if the bill is fake or genuine)

Variables

	Name	Description	Type	Units of Measurement
1.	is_genuine	If the bill if genuine or not	Bool	True/ False
2.	diagonal	The diagonal measurements of the bill	Quantitative	Millimeters (mm)
3.	height_left	The height of the left side of the bill	Quantitative	Millimeters (mm)
4.	height_right	The height of the right side of the bill	Quantitative	Millimeters (mm)
5.	margin_low	The lower margin of the bill	Quantitative	Millimeters (mm)
6.	margin_up	The upper margin of the bill	Quantitative	Millimeters (mm)
7.	length	The total length of the bill	Quantitative	Millimeters (mm)

Data Preparation

To be able to work and analyze the data, it needs to be prepared first. The head of the initial dataset containing the first 6 instances is shown as Figure 1 in the appendix. This step included multiple elements, beginning with converting the column *is_genuine* into a categorical variable with 2 classes: True and False.

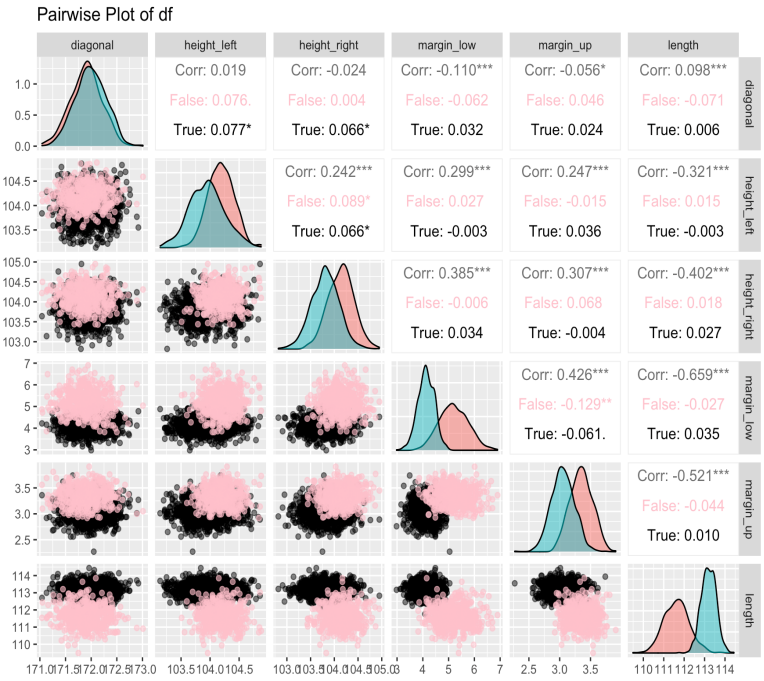
Next, the data was checked to see if it contained any missing values. It appeared that the dataset contained 37 missing values in total which were in 37 observations out of the 1500. All the missing values were in only 1 column: *margin_low*. Missing value imputation was performed for this continuous numeric column. The missing values were initially replaced by the median of the column, since it is a robust measure that is not affected by outliers, but since the data was seen to contain no outliers, as demonstrated in the following paragraphs, the missing values were changed to be replaced by the mean of the column instead of the median.

To analyze the dependency relationship between the different features in our dataset, the column rank was calculated. The column rank was 7, which is equal to the number of variables in the dataset, illustrating that the dataset is a full column rank matrix; all variables are linearly independent from each other.

The following step in the data preparation process is outlier detection. Since the data is multivariate with 7 variables, the blocked adaptive computationally efficient outlier nominators (BACON) algorithm proposed by Billor, Hadi, and Velleman (2000) was utilized to identify any outliers. This technique was applied only on numeric columns, which are all the columns except for the categorical *is_genuine* column (6 numeric features). The initial basic subset had 24 observations, and the computed threshold value was 5.544741. The basic subset was adjusted

according to this threshold, converging after 5 iterations; the final basic subset contained all 1500 observations hence the dataset contains no outliers. For confirmation, each group was checked for containing outliers separately, yet none of the 2 groups contained outliers as well.

After ensuring that the data contains no outliers, the relationship between the variables can be examined. First, the correlation coefficient matrix (figure 2 in the appendix) was studied. The 2 features with the strongest positive correlation were *margin_up* and *margin_low*. On the other hand, *length* and *margin_low* had the strongest negative correlation. In addition, the pairwise scatter plot of the variables, shown below, was also examined. It was coloured according to the label, *is_genuine* to be more informative. It appeared that in all pairwise scatter plots shown below, the 2 group of bills, both fake and genuine, were overlapping; it is also visible that all the variables follow a normal distribution but with different values for the mean and variance indicating different distributions although they are all measured in the same unit of measurement which is millimeter. Thus, there is no need to scale the data.



Discriminant Analysis

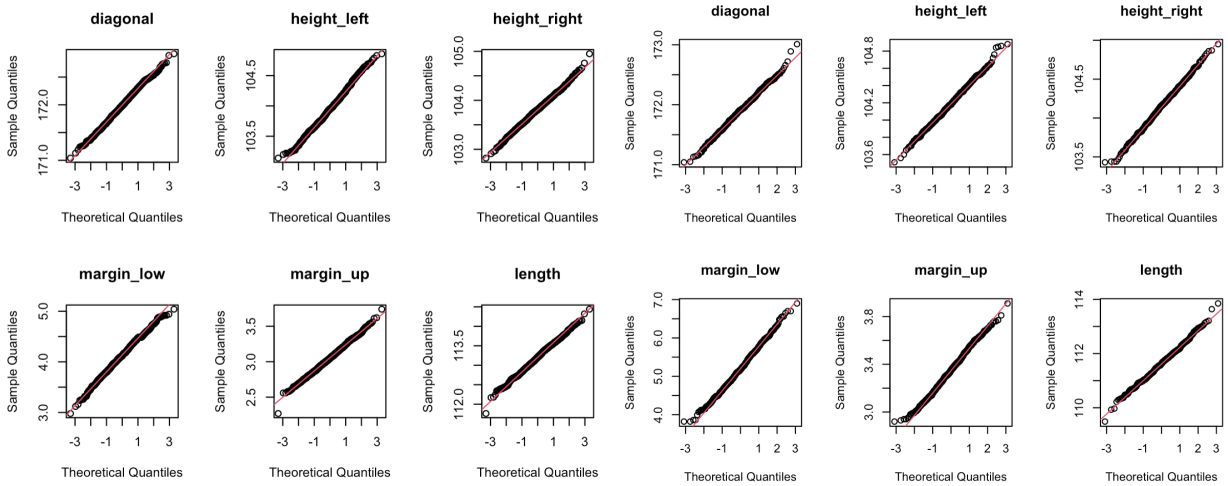
Since the data has been prepared, statistical analysis can be performed on it. The objective of discriminant analysis is to find a linear combination of a data's features that separate two or more classes within the data. As can be seen in the data description, the 'Fake Bills' dataset consists of two labels or classes: '1' referring to genuine bills and '0' referring to fake ones. The data was standardized and a pairwise plot was created to show the two classes with homogenized scaling. From the resulting plot, it is visible from the compact nature of each cluster that each of the two groups has low within-class dispersion (S_W). The two groups were, however, overlapping in all of the pairwise plots, indicating low between class dispersion (S_B).

Fisher's Linear Discriminant Analysis (FLDA)

In order to perform Fisher's Linear Discriminant Analysis, two assumptions must be verified. First, the data must be shown to be approximately normal. According to the central limit theorem, this assumption should be true since the number of observations in each group is relatively large (exceeds 30). The second assumption is that the covariance matrices of the data must be identical across the data's classes; in other words, the covariance matrices of both classes should be approximately the same. This assumption will not be verified and will be assumed to be true.

To check for normality, the pairwise scatter plots containing the graphs of the two groups (genuine and fake bills) were examined for each of the data's variables. As clear from the plots above, the graphs show that each of the variables is represented by a bell-shaped curve, thus visually supporting the assumption of normality for the data. Furthermore, the normal QQ-Plots for each label (shown below) were studied, showing that the vast majority of the corresponding

points fall on the trend line. This further supports the normality assumptions, so this assumption has been verified. Continuing with the assumption of equality of covariances, Fisher's Linear Discriminant Analysis can be implemented.

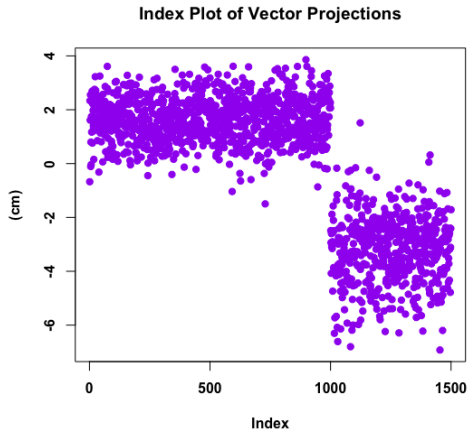


First, the data's variables were separated from the label. The discriminant analysis function was trained on the data and then used to predict the label of each observation. Figure 3.1 in the appendix shows the table generated to internally validate the resulting classifications showing the correctly classified proportion of the data as well as the misclassifications. The error rate of internal validation was approximately 1.067%, 14 false positives and 2 false negatives. This is a relatively small error rate thus the classification was accurately performed.

Next, external validation was performed using Leave-One-Out cross validation. The output and error rate was found to be identical to that of the internally validated classification: 1.067%. This is likely due to the fact that the two clusters are very clearly defined. Figure 3.2 in the appendix shows the confusion matrix generated from the external classification.

In order to evaluate the vector projections of the resulting classification, an index plot was created (depicted below) showing the LDA function's classification scores. From the plot, it

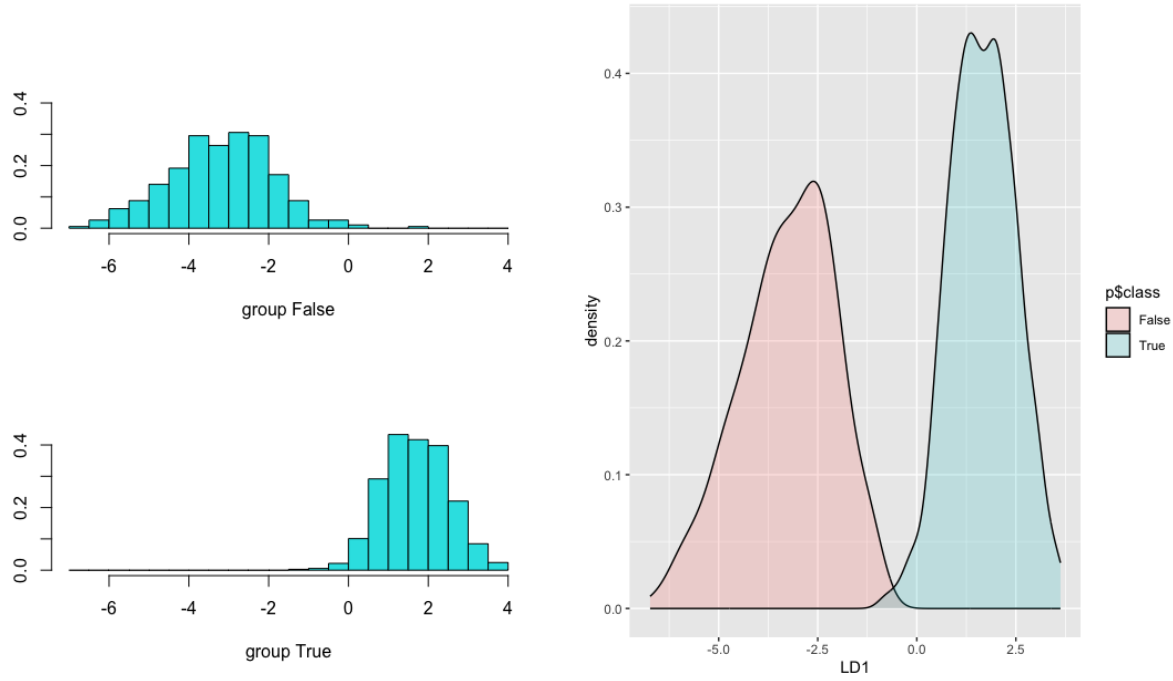
is apparent that two general clusters are formed; The first cluster is found to consist of positive values and is more frequently occurring compared to the second cluster, which consists of negative values. The positive values (first cluster) refer to the genuine dollar bills (*is_genuine* = True), while the negative values (second cluster) refer to the counterfeit bills. While two clear groups are formed in the plot, a few outliers can also be spotted which indicates that there will be a few misclassifications. However, since an overwhelming majority of the points lie within one of the two groups, it is expected that the misclassification rate will be low (which was observed from the low error rates from the validation of the classification results in the previous section).



As a way of experimenting with alternative methods of external validation, the data was split into two groups, one for training (75% of the data) and one for testing (25% of the data). This will create a fixed testing set over which external validation of the prediction model can be executed. Next, linear discriminant analysis ('lda') was performed on the training data, and the model's output included information on the prior probabilities of the two classes, the group means of each of the two classes, as well as the coefficients of the linear discriminants. The genuine bills were shown to have a prior probability (proportion of the training set) of 0.655

while the fake bills had a prior probability of 0.344. The column means of the fake bills class within each feature in the training set were generally found to be slightly higher than those of the genuine bills class except the variables *length* and *diagonal*, which were found to be, on average, slightly higher among the genuine bills class. The coefficients of the linear discriminants (the linear combination of the variables used for discriminating between the classes) were also listed, with the largest contributor being the *length* of the dollar bill. The 2 plots below demonstrate the predicted classification results.

The plots show that there is very little overlapping between the two classes, which means that there is a higher likelihood of accurately classifying a dollar bill given the data's features. The results also demonstrate that the fake dollar bills, classified under the group 'False' are more widely dispersed than the genuine bills. This is expected since genuine dollar bills have fixed dimensions that allow for little to no variation. In addition, the classifications of genuine dollar bills were more frequent in the training data than that of counterfeit bills. The error rate of the training set (internal validation) was 0.98214%; for the testing set (external validation), the error rate was 1.0526%. The exact number of observations classified in each class are illustrated in the confusion matrices in figures 3.3 and 3.4 in the appendix. These results indicate that Fisher's discriminant was, in fact, able to correctly classify the vast majority of observations in the training set.



FLDA-2

In addition to performing Fisher's Linear Discriminant Analysis on all the features in the dataset to classify each observation into one of the 2 groups, FLDA-2 was also implemented. This attempts to classify the observations into the 2 groups based on only 2 variables from the 5 in the dataset hence facilitating visualizing the separability of the data since a two dimensional plot can be analyzed. The 2 variables chosen are the ones who's plot has the highest separability between the 2 groups, or in other words, the 2 groups appear distinct the most in the plot of these 2 variables against each other. Using the pairwise plot of the df shown above, 3 pairs of variables were selected, allowing FLDA 2 to be implemented 3 times. First, the plot of the variables *margin_low* and *length* showed the least overlap between the 2 groups compared to the others, and indeed, the error rate when Fisher's Linear Discriminant Analysis was done using these

variables yielded an error rate of 1.93% due to 29 misclassifications. Figure 4.1 in the appendix shows the threshold (line separating the two classes) having a positive slope, meaning that the variables have a positive relationship and it is not orthogonal to the line connecting the means, which indicates a low S_B to S_W ratio yet the low error rate indicates a high between-class dispersion. The non-orthogonal nature of these two lines also points to unequal covariance matrices between the two classes in the data.

The variables *margin_up* and *length* also seemed appropriate for implementing FLDA 2 according to their pairwise plot. The error rate using these variables was higher than the previous pair; it was 3.07% due to 46 misclassifications. Similar to the plot of the previous pair of variables, the projection vector separating the 2 groups has a positive slope as seen from figure 4.2 and it is not orthogonal to the line connecting the means, which indicates a low S_B to S_W ratio, as well as unequal covariance matrices between the classes. For the final pair of variables, *height_right* and *length* were used. This pair had the highest error rate, 3.6% due to 54 misclassifications, out of all 3 pairs of variables. From merely observing the pairwise scatterplot of these two variables, one can deduce that the error rate will be relatively high due to the high overlap of points compared to the previous variables. The relationship of these variables is also positive as indicated by the slope of the projection vector in Figure 4.3 and there is a low S_B to S_W ratio as seen from the non-orthogonal line connecting the means and the projection vector. Thus, it can be concluded that *margin_low* and *length* are the best 2 variables to select for classification if only 2 variables will be used.

Multinomial Distribution: An Alternative Discriminant Analysis Method

Lastly, the data was classified using the multinomial distribution method. The results show that error rate is 0.87%; there were less occurrences of misclassification using this method (13 in total) than there were using Fisher's discriminant (16 in total). Nevertheless, the difference is insignificant as it only accounts for 0.2% of the data. The confusion matrix for this method can be found in the appendix as Figure 5.1. For external validation, Leave-One-Out cross validation yielded an error rate slightly higher, which is 0.93% due to 14 misclassifications, depicted in figure 5.2 in the appendix.

Cluster Analysis

In cluster analysis, the data consisting of 1500 observations is divided into clusters based on the features for each observation excluding the *is_genuine* variable. The goal is to minimize the within-group variation and maximize the between-group variation of the clusters. Because cluster analysis is an unsupervised learning method where the data is split into clusters that are not previously known, the column *is_genuine* is dropped from the data frame and is stored separately, only for validation. Now the number of clusters in the dataset is assumed to be unknown; Since the data is not two-dimensional, the clusters cannot be visually determined via plotting the points, so multiple techniques will be tried to best split the data into clusters (which is already known to be 2).

Hierarchical Algorithm (Agglomerative)

The first technique used for clustering is the agglomerative hierarchical algorithm. In this technique, a dendrogram is created based on splitting the data from n clusters (1500 in this case) to a single cluster, illustrating the distances in each case. The number of clusters is determined

based on the greatest distance between the clusters represented by the longest vertical lines in the dendrogram. There are multiple techniques to measure the intercluster and intracluster distances, which are the distances between the different clusters and distances between the points within the same cluster respectively.

There were 6 different combinations of the between observations and between cluster distances implemented, each in a separate iteration on the data, to select the best clustering technique. For the distances between observations, the Euclidean distance and the Manhattan distance were the measures used since all the features are numeric and there are no categorical features in the dataset. For the distances between clusters, Ward's distance, average linkage, and distance between cluster centroids were used. To evaluate each method of clustering, the error rate, which is a measure of misclassified observations, was used; the method that has the lowest error rate, which is a percentage ranging from 0 to 100, is the best. It is worth noting that the R^2 was also calculated, but all the methods yielded a perfect R^2 (equal to 1), indicating that the dataset itself is not of low quality, yet it does not necessarily mean that the clustering is good. The dendrograms for all the methods are shown in the appendix (figures 6.1 to 6.6).

The best method used Euclidean distance as the measure of distance between points and Ward's distance for the distance between clusters. This method was the most suitable to the dataset as it had the lowest error rate, 0.933 %, out of all the 6 methods, and misclassified 14 out of 1500 observations. In addition, the dendrogram, figure 3.1 in the appendix, showed that the data is clearly separable into 2 main clusters. Likewise, the second method depending on Manhattan distance and Ward's distance also had a slightly higher yet low error rate of 1.07% due to 16 misclassifications. Methods 3 and 4, using Euclidean distance and average linkage; and

Manhattan distance and average linkage, respectively, had very similar results; their error rates for method 3 was 1.13% with 17 misclassifications and 1.2% for method 4 due to 18 misclassifications. This indicates that their performance was worse than the first 2 methods although still not very poor. In all the aforementioned methods, there were more false negatives than false positives since the number of observations that are *True* are more than *False* ones.

As for the last 2 methods, the measures of distance used were unsuitable to the dataset because they have underperformed. When the Euclidean distance and cluster centroids were used, the error rate was 33.27%, which is very high. All the points were clustered together except for a single observation, as visible from the dendrogram. After studying this observation, it appeared that it is a fake bill, *is_genuine* = False, which is logical since authentic bills have almost identical measurements. As for method 6 which depended on the Manhattan distance and cluster centroids, the error rate was 33.2% because only 2 points were considered in cluster 2 and the rest in cluster 1. As expected, these 2 points appeared to be fake bills and they were different from the observation isolated by method 5.

One conclusion that can be drawn from these results is that these 2 observations of cluster 2 in method 6, or the single observation distinguished by method 5, can be considered outliers, although BACON did not identify them, or have unusual features. The index plots of all the variables were examined and this single point in method 6 was marked with a red cross to understand why it appeared as an outlier. Nevertheless, there was no distinct feature where this observation appeared to vary from the rest of the observations as shown in figure 7.1 in the appendix, and likewise for the 2 observations isolated in method 6 (index plots shown in figure

7.2 in the appendix). This suggests that this single observation is not an outlier in the 1-dimensional space but is an outlier in the 5-dimensional space (all the variables together).

k-Means

The k-Means clustering algorithm was also applied on the dataset. This algorithm is relatively simple and quick; it fits the dataset as its mean is defined due to it being numeric. It requires the number of clusters (k) to be specified in advance, which in this case is known to be 2. The R function used depended on the Euclidean distance as the default measure of distance between points. The R^2 was equal to 1, as all the previous methods, and the error rate was equal to 1.67% due to 25 misclassified observations. Relative to the first 4 methods of hierarchical clustering, this method underperformed. To ensure that 2 is the best number of clusters that the data can be divided into, the L-curve (figure 8 in the appendix) was also plotted for values of k from 1 to 10, supporting the assumption that the data is naturally divided into 2 groups. This L-curve is shown below.

Conclusion

In conclusion, the aim of this project was to apply the multivariate analysis technique: discriminant analysis and cluster analysis, as well as other methods during data preparation, in order to distinguish genuine mention from its counterfeit counterpart thus allowing for better identification of fake dollar bills. This was done use the dataset (fakebills.csv) which consisted of 1500 bills and 6 measurements for each as to whether the bill is authentic or not. When Fisher's Linear Discriminant Analysis was applied on all the variables in the dataset, after verifying the assumptions, on only 2 features in the dataset, and multinomial distribution, it appeared that the first technique was the best to distinguish between the 2 groups in the dataset as it was the most accurate with the lowest error rate. For cluster analysis, agglomerative hierarchical clustering using 6 different combinations of intercluster and intracluster distances and k-Means were used to find the best way of separating the clusters in the dataset. The L-cure of k-Means also indicated that the dataset indeed should be divided into 2 clusters, and the method that yielded the lowest error rate was hierarchical clustering depended on Euclidean distance for measuring the distance between points and Ward's distance for the distance between clusters. 3 observations stood out during cluster analysis during different methods, suggesting that they require further investigation as to whether they are outliers or not despite not being detected by BACON.

Appendix

Figure 1: Head of initial Dataset

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
1	True	171.81	104.86	104.95	4.52	2.89	112.83
2	True	171.46	103.36	103.66	3.77	2.99	113.09
3	True	172.69	104.48	103.50	4.40	2.94	113.16
4	True	171.36	103.91	103.94	3.62	3.01	113.51
5	True	171.73	104.28	103.46	4.04	3.48	112.54
6	True	172.17	103.74	104.08	4.42	2.95	112.81

Figure 2: Correlation Coefficient Matrix

```
              diagonal height_left height_right margin_low   margin_up      length
diagonal      1.00000000  0.01947232  -0.02449201 -0.1102436 -0.05564888  0.09758729
height_left   0.01947232  1.00000000   0.24227881  0.2990280  0.24652224 -0.32086276
height_right -0.02449201  0.24227881   1.00000000  0.3845134  0.30700464 -0.40175122
margin_low   -0.11024358  0.29902799   0.38451340  1.0000000  0.42560768 -0.65884370
margin_up    -0.05564888  0.24652224   0.30700464  0.4256077  1.00000000 -0.52057513
length        0.09758729 -0.32086276  -0.40175122 -0.6588437 -0.52057513  1.00000000
```

Figure 3.1: FLDA Internal Validation Output

Predicted	**Actual**	
	False (Counterfeit Bills)	True (Genuine Bills)
False (Counterfeit Bills)	486	14
True (Genuine Bills)	2	998

Figure 3.2: FLDA External Validation Output

Predicted	**Actual**	
	False (Counterfeit Bills)	True (Genuine Bills)
False (Counterfeit Bills)	486	14
True (Genuine Bills)	2	998

Figure 3.3: Internal Validation of Training Set (75% of data)

Predicted	**Actual**	
	False (Counterfeit Bills)	True (Genuine Bills)
False (Counterfeit Bills)	376	1
True (Genuine Bills)	10	733

Figure 3.4: External Validation using Testing Set (25% of data)

Predicted	**Actual**	
	False (Counterfeit Bills)	True (Genuine Bills)
False (Counterfeit Bills)	111	1
True (Genuine Bills)	3	265

Figure 4.1: FLDA 2 using *margin_low* and *length*

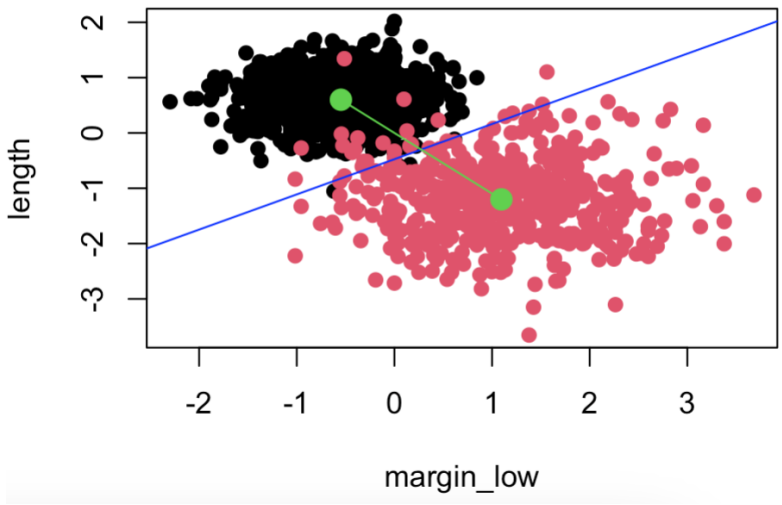


Figure 4.2: FLDA 2 using *margin_up* and *length*

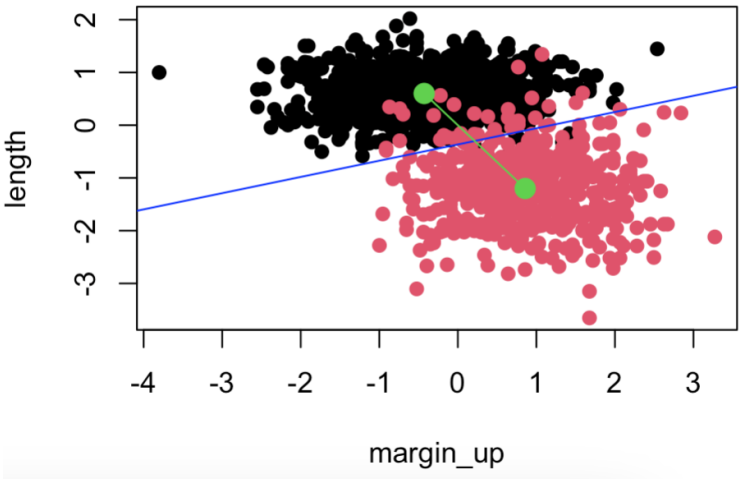


Figure 4.2: FLDA 2 using *height_right* and *length*

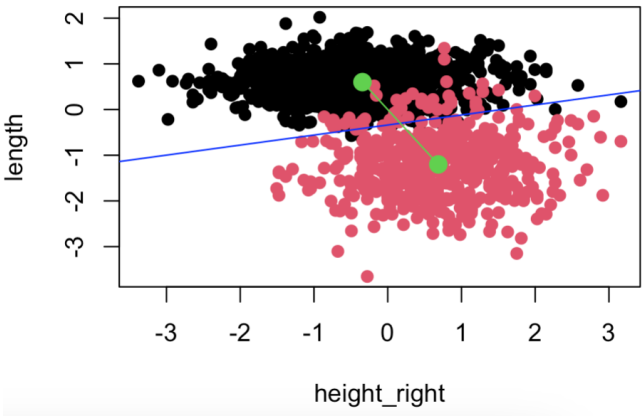


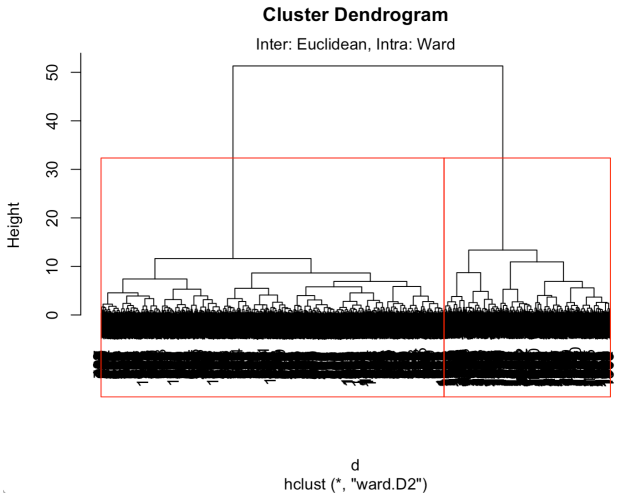
Figure 5.1: Multinomial Discriminant Analysis Internal Validation

Predicted	**Actual**	
	False (Counterfeit Bills)	True (Genuine Bills)
False (Counterfeit Bills)	491	9
True (Genuine Bills)	4	996

Figure 5.2: Multinomial Discriminant Analysis External Validation

Predicted	Actual	
	False (Counterfeit Bills)	True (Genuine Bills)
False (Counterfeit Bills)	491	9
True (Genuine Bills)	5	995

Figures 6.1 and 6.2: Dendrogram of Hierarchical Clustering method 1 and 2



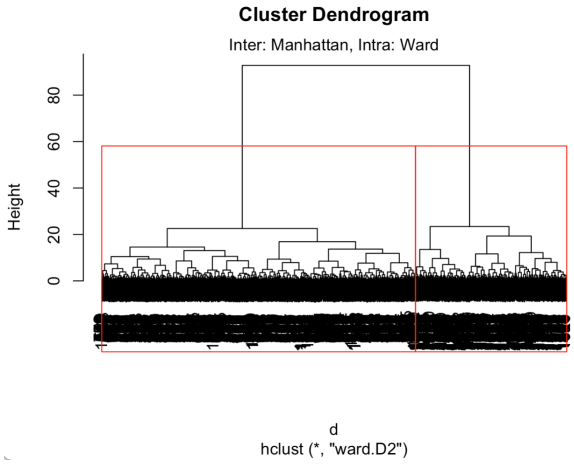
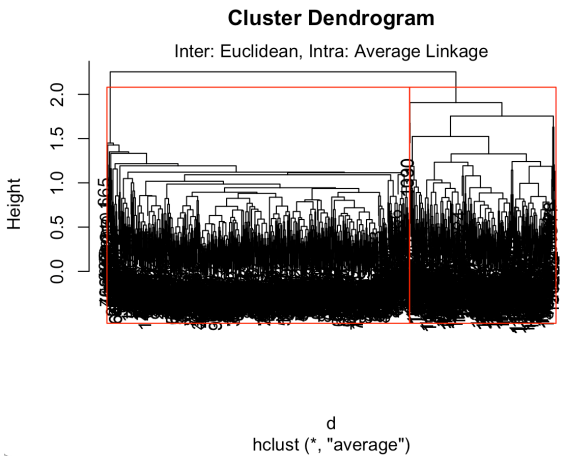


Figure 6.3 and 6.4: Dendrogram of Hierarchical

Figures 6.3 and 6.4: Dendrogram of Hierarchical Clustering method 3 and 4



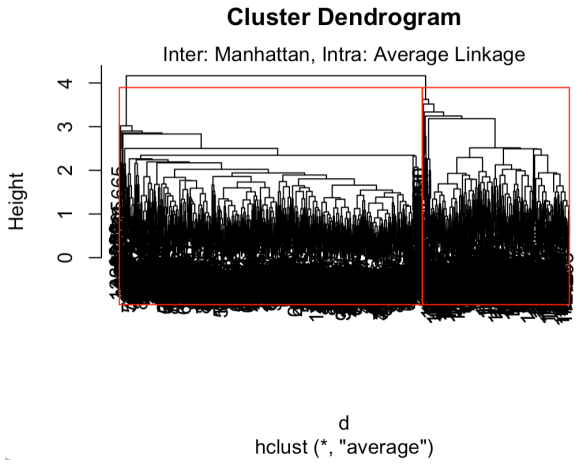


Figure 6.5 and 6.6 : Dendrogram of Hierarchical Clustering method 5 and method 6

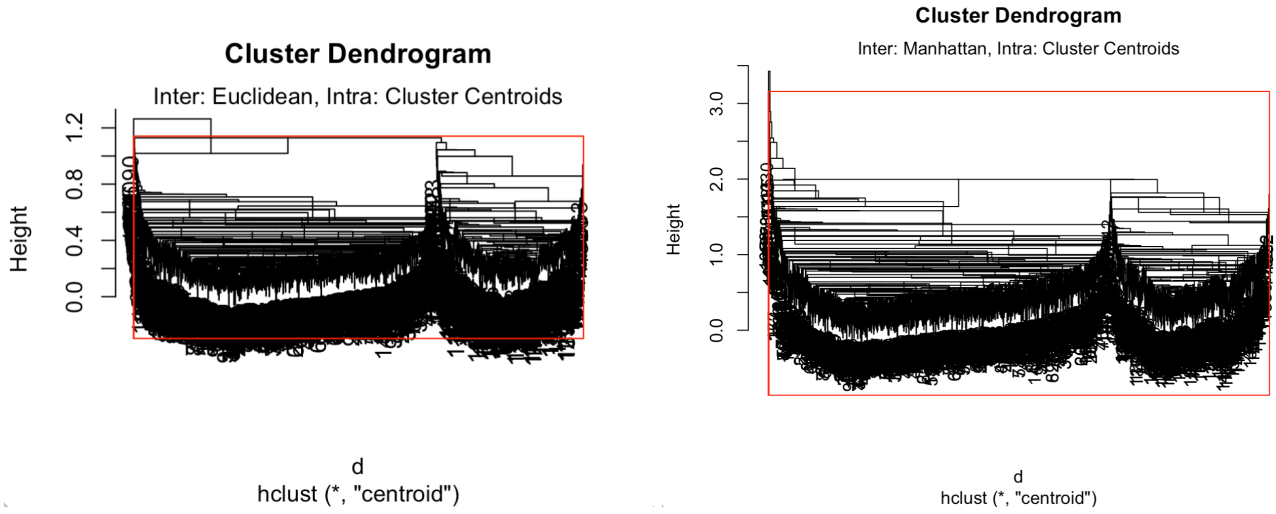


Figure 7.1: Index plot of features in Hierarchical Clustering method 5

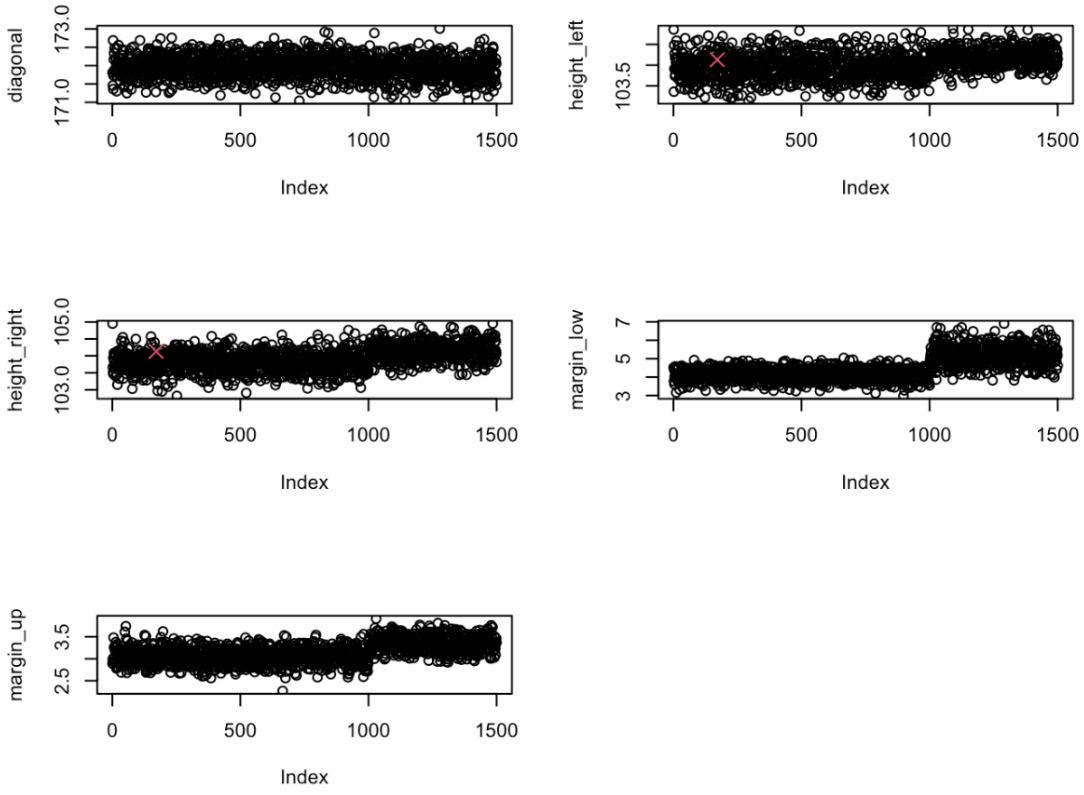


Figure 7.2: Index plot of features in Hierarchical Clustering method 6

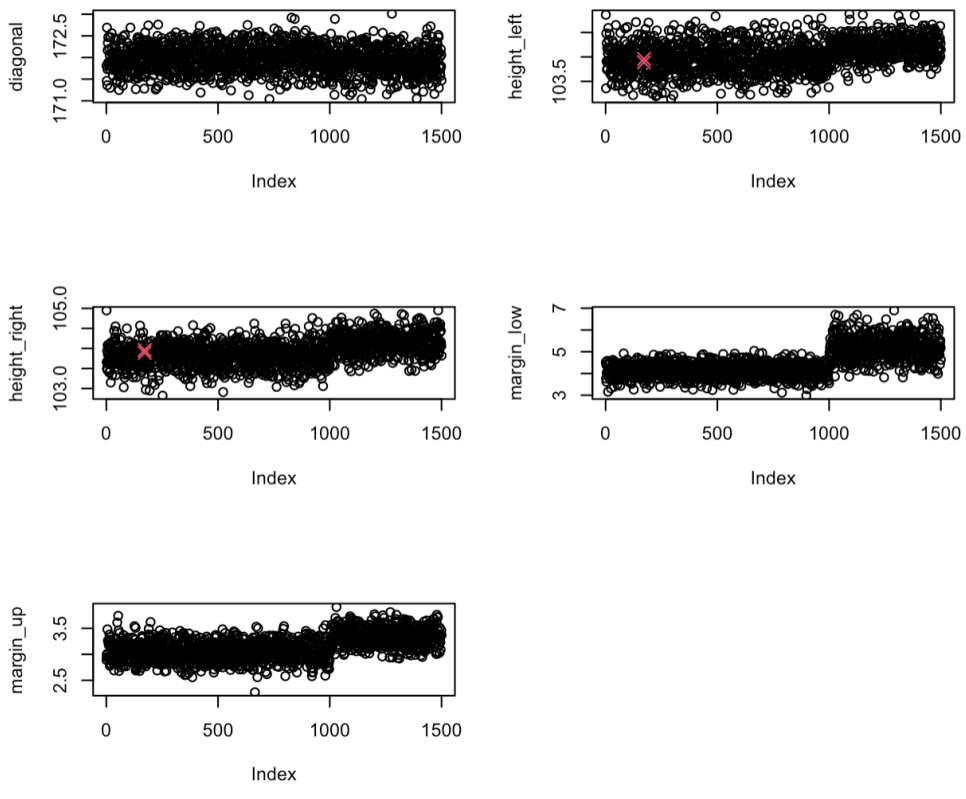


Figure 8: The L-Curve for K-means clustering