University of Birmingham

Intelligent Data Analysis Assignment

Nik Zulhilmi Nik Fuaad

1446348

Spring Term 2018

Abstract

The purpose of this work was to explore the principal component analysis (PCA) techniques and coordinate projection in analysing a set of data, find the relationships between varying attributes, and determine the predicted attributes based on the rest of the attributes. Although PCA is more commonly used for analysing huge set of data with tens, hundreds, and more attributes, it still possesses some significant advantages. The data set going to be used is about Iris plant and its physical characteristics.

Contents

1. Introduction

Iris is a flowering plant which belongs to the genus of 260-300. Some biologists stated the name refers to the huge variety of flower colours and sizes among the many species [1]. The plant has interesting morphologic variations. It was the purpose of this project to investigate the relationships between the morphologic variation and the species it belongs to.

The morphologic variations going to be considered are the sepals and petals of the flowers. Sepal is the outermost circle of a flower part that encloses a bud before it opens. Petals are the modified leaves which are inside of sepals. The species considered in this project are Iris Setosa, Iris Versicolor, and Iris Virginica.

2. Pre-processing of the Data

The data set contains 150 instances and consists of 4 attributes and 3 classes, 50 for each class. The attributes are sepal length, sepal width, petal length, and petal width, all in centimetres (cm). The classes are the species mentioned above in the introduction.

2.1 Labelling

Since the data is already labelled into three classes, no further classification or labelling was done. The data was ordered according to the classes where the first 50 instances belong to the class Iris Setosa, the following 50 belong to Iris Versicolor, and the remaining instances belong to Iris Virginica. The original data file was cleaned to contain only numerical features and replaced the class names with numbers. Numbers 1,2, and 3 to indicate the data belongs to classes Iris Setosa, Iris Versicolor, and Iris Virginica respectively. The processed data can be found in irisdata.txt file attached.

Figure 1 and Figure 2 show the distribution, central value, and the variability of each attribute for each class. The measurements in the data are scaled properly and no further scaling is needed. The unit used are consistent for all attributes. Hence, conversion to different units is not required. Besides, since the number of instances is only 150, it is decided not to quantify the data into different intervals. For example: short, medium, long, and etcetera.

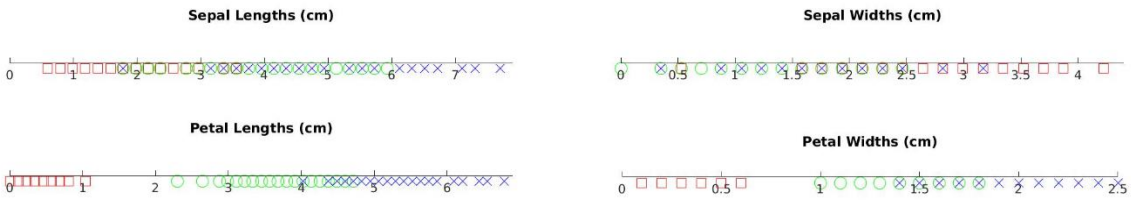

Figure 1: Line plots showing each of the attributes for the three classes. Class 1 is red, class 2 is green, and class 3 is blue.

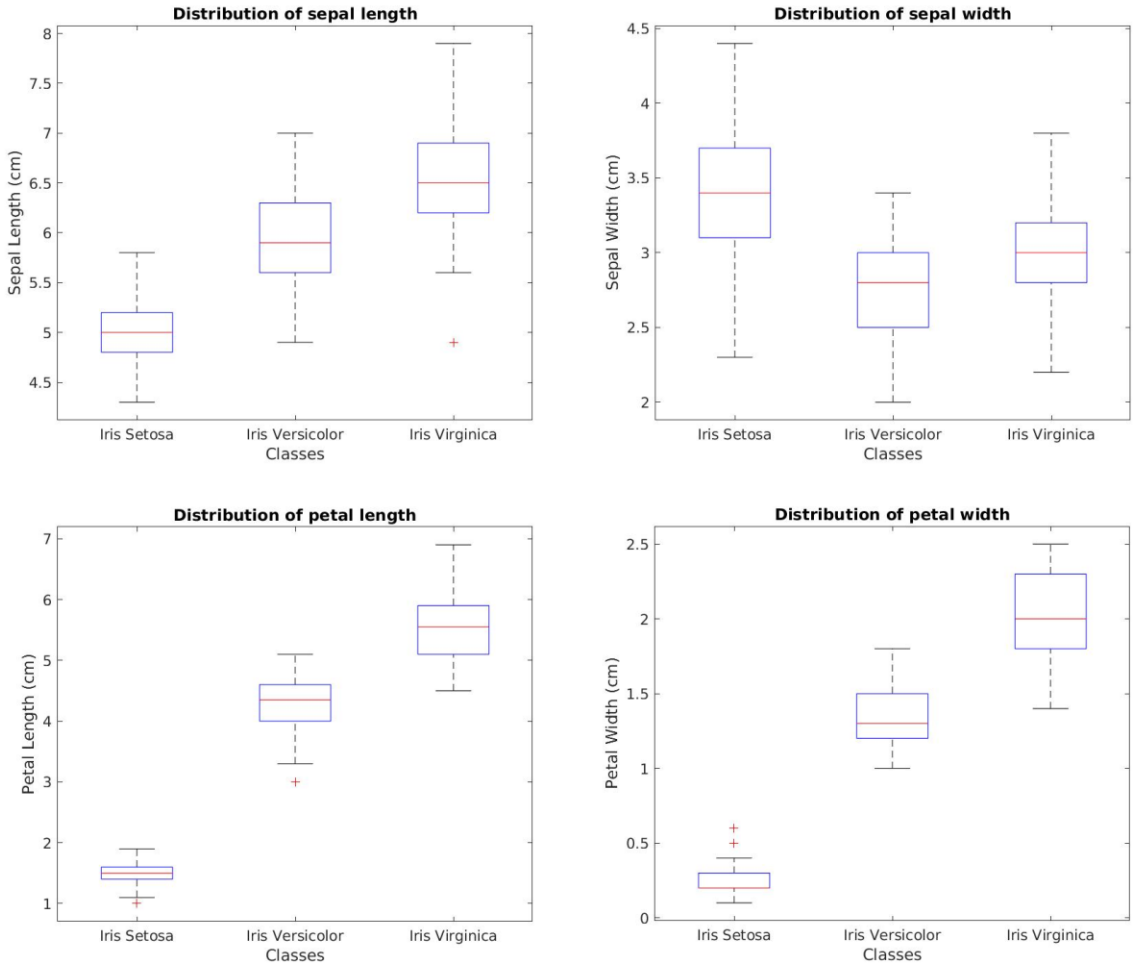

Figure 2: Box plots showing the range of values for each of the attributes.

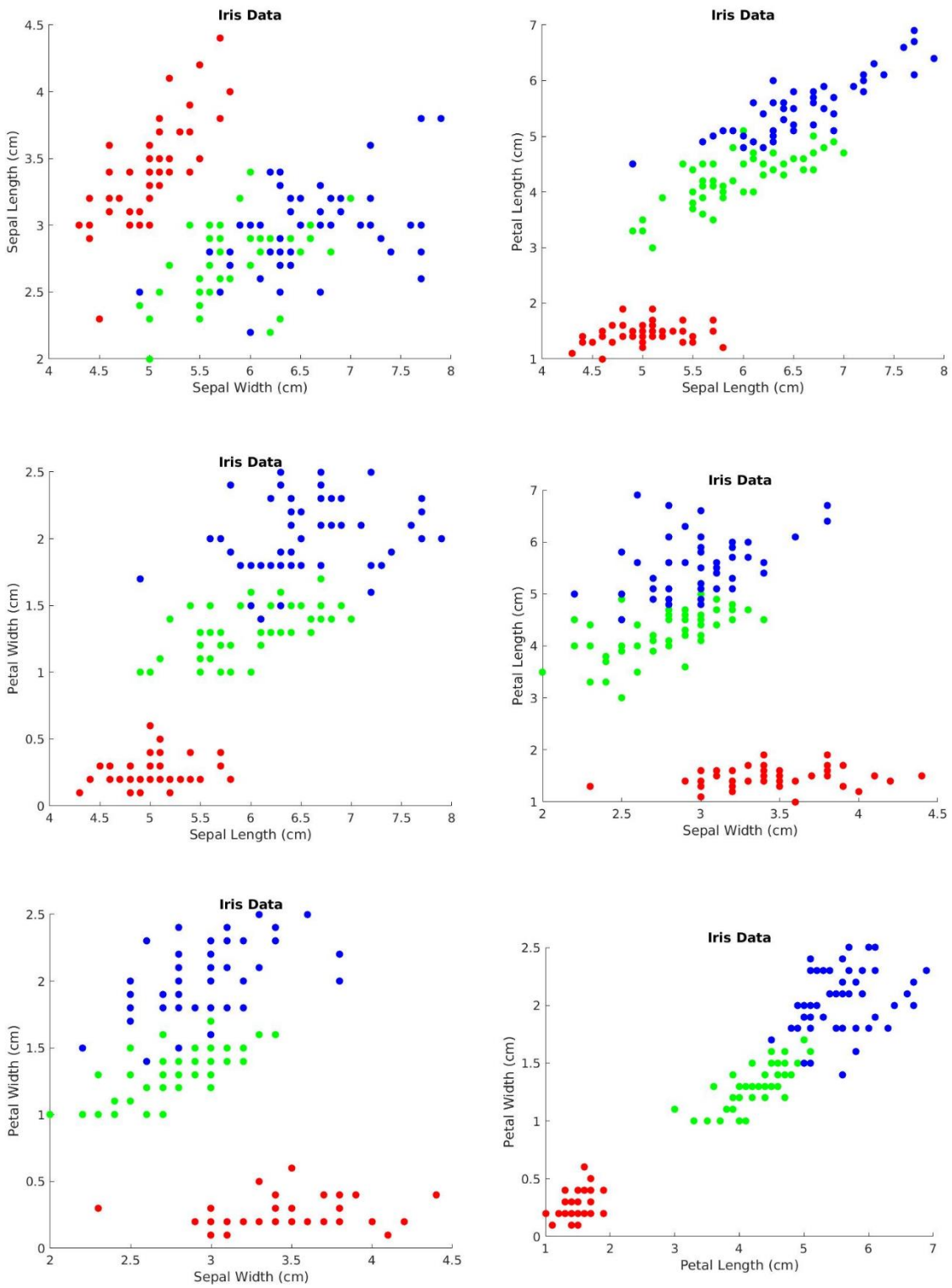

Figure 3: 2-D plots showing the relationship of each attributes with each other. Class 1 is red, class 2 is green, and class 3 is blue.

3. Principal Component Analysis

Before principal component analysis is done, each of the attributes were compared against each other for all the classes. The objective was to visualise the data using 2-D plots and identify what are the most important axes in determining the classes of the plant. In Figure 1 and Figure 2, intuitively, it could be argued that the classes can be distinguished from some of the attributes, namely petal length and petal width. But those two attributes don't provide clear distinction. In Figure 3, the distinction between the classes are clear in all the comparisons with only small overlapping, except for sepal width vs sepal length. Figure 3 was generated with reference to [2].

From the 2-D plots, class 3 is easily identified where the clusters are further away than the other two classes. For class 2 and class 3, the clusters formed only overlap a little.

One of the disadvantages of the 2-D plot is it is not suitable be used to visualise data with more than 4 dimensions as a lot of comparisons need to be done. After the above data has been compared and briefly analysed by plotting the relationships between the attributes of each classes, MATLAB's functions [3] were used to perform principal component analysis to generate the eigenvalues, eigenvectors, covariance matrix, and coefficients to the eigenvalues. Principal component analysis is effective in analysing large data sets by using data dimension reduction techniques [4]. The dimensions are reduced while keeping as much variation as possible.

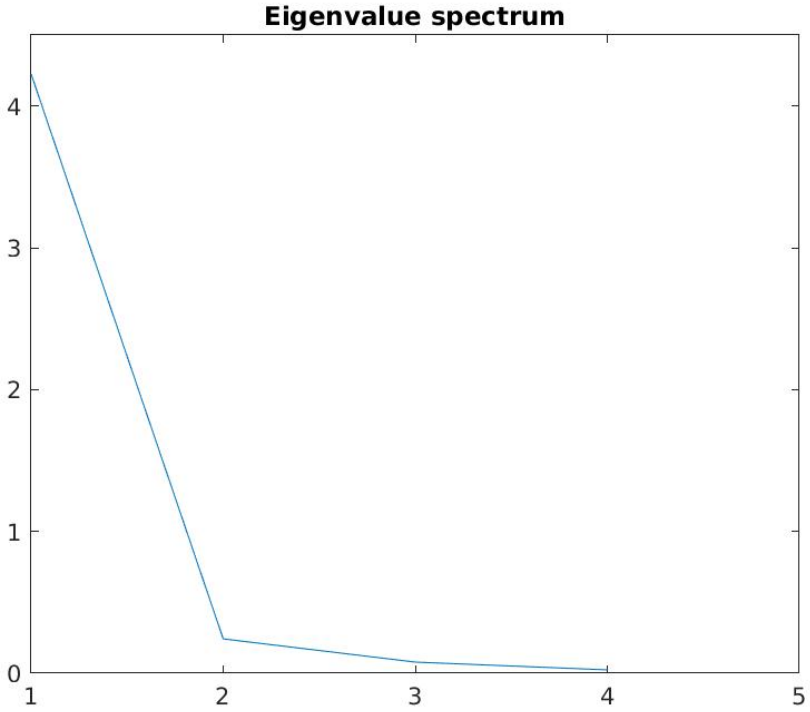

Figure 4: Eigenvalue spectrum shown in descending order.

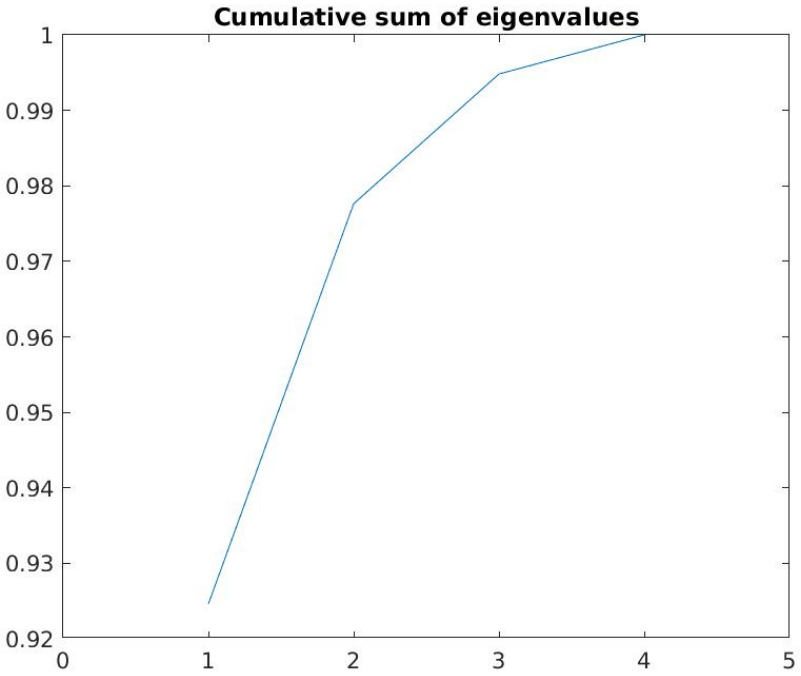

Figure 5: Cumulative sum of the eigenvalues in descending order.

3.1 Eigenvectors and Eigenvalues Analysis

The eigenvalues of the covariance matrix were plotted as shown in Figure 4 in descending order. The most data variance is explained by the first eigenvector of the covariance matrix. The first 2 eigenvectors explain almost all the variance in the data. Figure 5 shows the cumulative sum of all the eigenvalues in descending order. The first eigenvector explains 92.5% of the variance the data. The first two eigenvectors explain a total of 97.8% of the variance of the data and will not lead to losing much information if they are used in the projection of the principal component analysis component space.

Once the main principal components were identified, a 2-D plot was generated and can be seen in Figure 6. The projection in Figure 6 shows a good finding where the clusters or each class are easily identified. The result could be improved even more by considering the third eigenvector to generate a 3-D plot with three principal components. The first three eigenvectors capture a total of 99.5% of the variance of the data, which is just 1.7% more than using only the first two. Hence, no 3-D plot was generated.

Subsequent to the 2-D plot, a biplot was generated to visualise the coefficients of the eigenvalues for all the attributes and the principal component scores for each data point. The biplot can be seen in Figure 7. The magnitude and the direction of the vectors generated indicates the contribution to the principal components from each attribute. Sepal length, petal length, and petal width contribute largely to the first principal component. Meanwhile, sepal length and sepal width contribute largely to the second principal component. Sepal length contributes to both principal components and it is deemed to play the most important role in classification of the plants

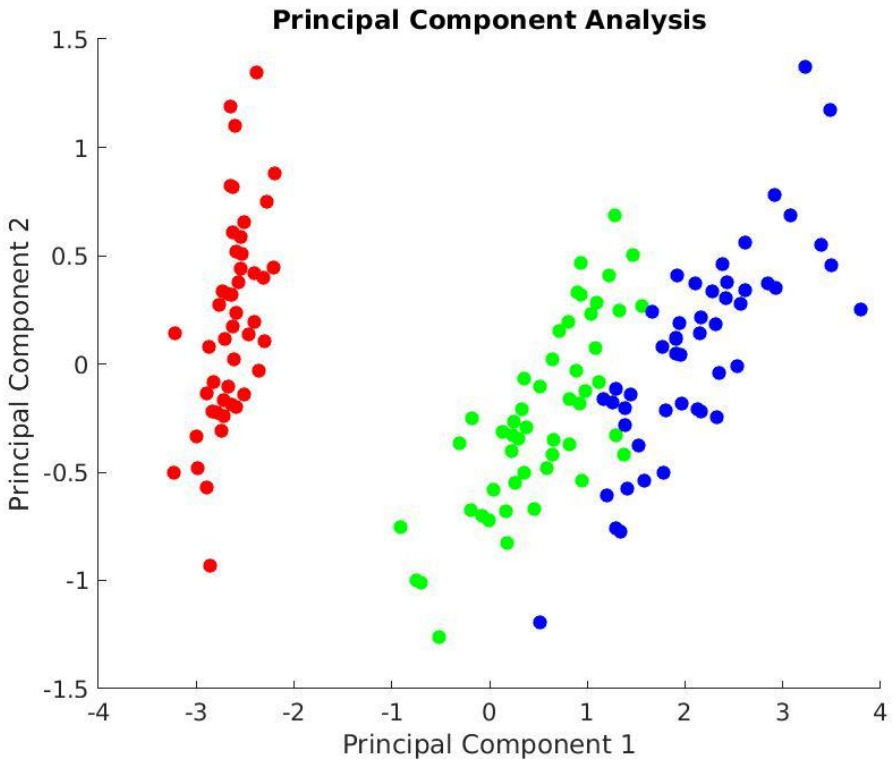

Figure 6: 2-D plot showing the projection of the principal component analysis results. Class 1 is red, class 2 is green, and class 3 is blue.

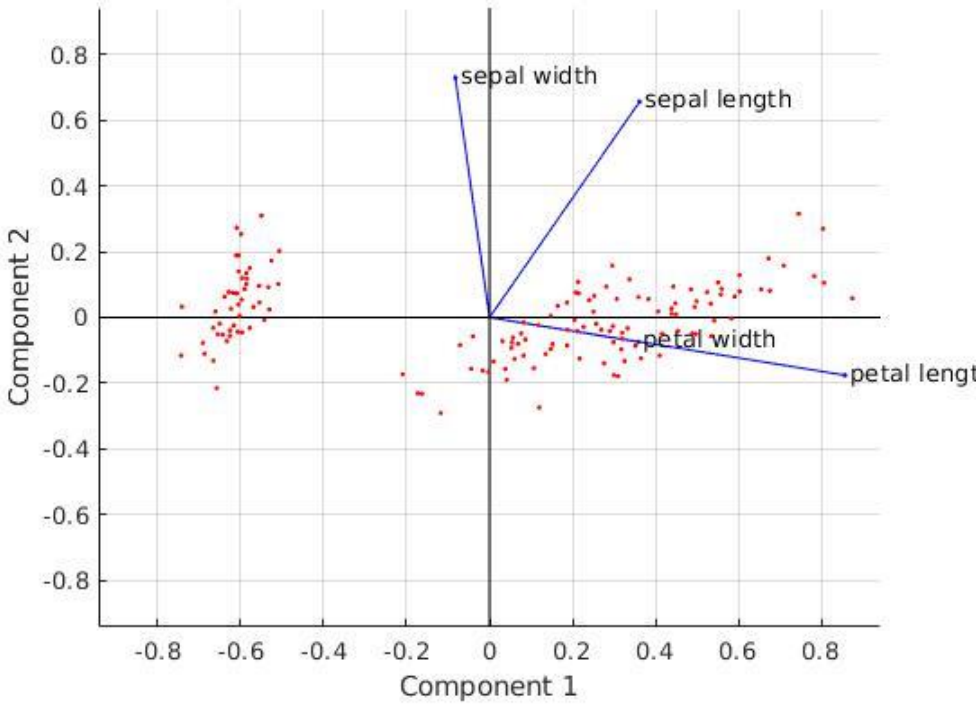

Figure 7: Biplot to visualise the coefficients of the eigenvalues and principal component scores for each data point.

Apart from that, vectors that point to directions close to each other shows similarity and relationships. It proves that petal width and petal length are strongly related. This is confirmed on the 2-D plot in Figure 3 in the comparison between petal width and petal length. The relationship shows a linear trend.

3.2 Discussion

Although one of the disadvantages of principal component analysis is the lost of data, it can be overcome by taking into consideration of more eigenvectors. This dimension reduction technique is useful in analysing large data set which can contain more than 50, or even hundreds of dimensions. Reducing the dimension allows the less significant and less dominant data to be neglected.

4 Further Pre-Processing of the Data

The Iris data set given is quite straightforward even without the use of principal component analysis because it contains only 4 attributes and 3 classes. One interesting this to consider is the size, instead of the length and width of the petals or sepals. In this section, it is going to be demonstrated the relationship between the petal sizes and sepal sizes of the flowers. Since the length and width is given, the size can be calculated by multiplying the two values. Although in reality, the size of a petal or sepal is not as straightforward and for the sake of simplicity and further analysis of the data, the aforementioned formula is going to be used. The sizes of petals and sepals are added as one dimension each into the data set.

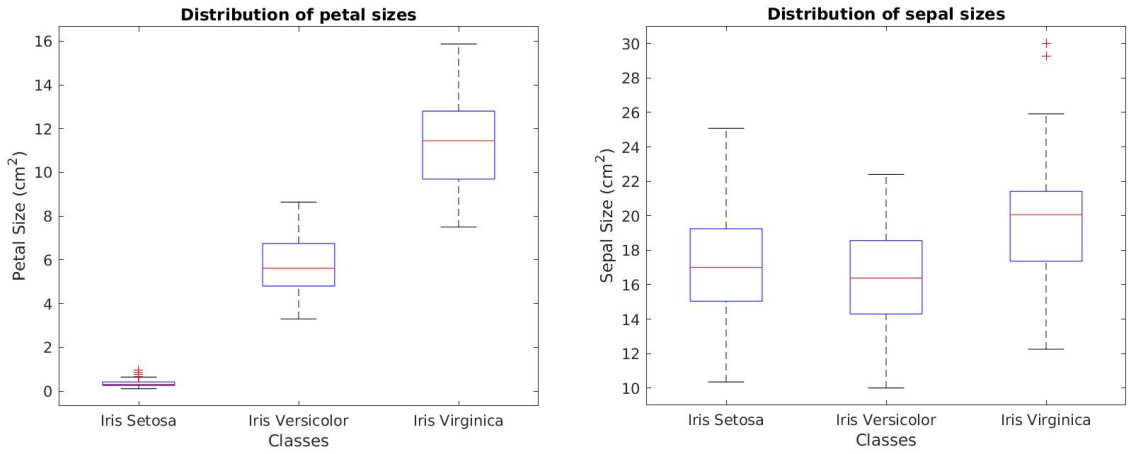

Figure 8: Box plots showing the distribution of petal sizes and sepal sizes for all classes.

In the distribution of the petal sizes in Figure 8, the distinctions are clear between the classes, especially for the class Iris Setosa. Again, since the sizes are derived from the previous attributes which already use consistent units and are properly scaled, the new dimensions are not required to perform conversions or scaling to the measurements.

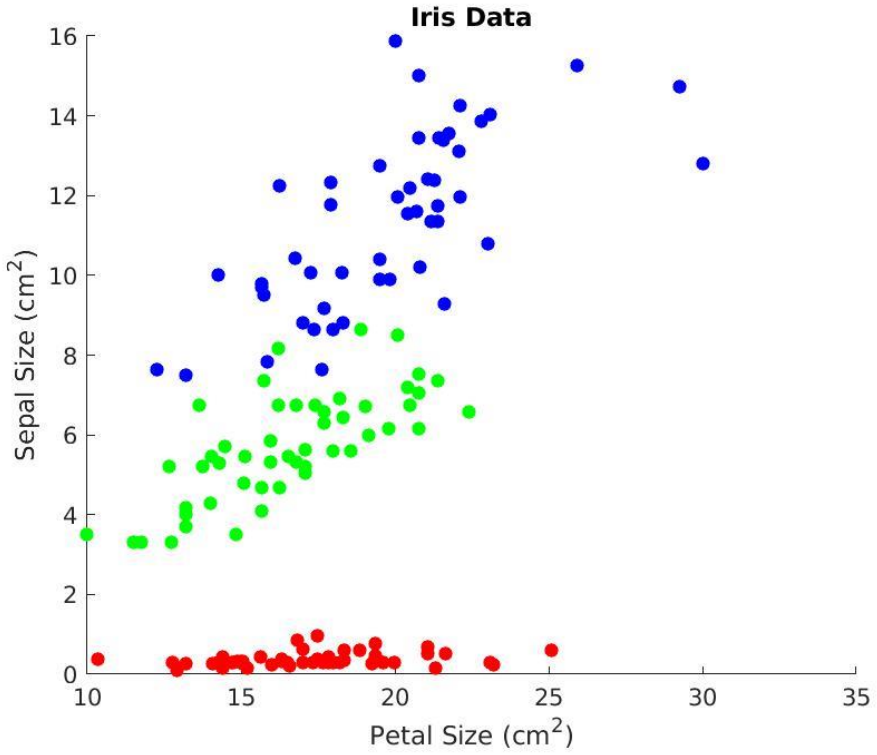

Figure 9: 2-D plot showing the relationship between petal size and sepal size for all classes. Class 1 is red, class 2 is green, and class 3 is blue.

In Figure 9, all clusters are easily identifiable as there are only a few points overlapping between the clusters of class 2 and class 3. Classes 2 and 3 show a linear trend where the size of the sepal increases as the petal size increases. For class 1, the sepal size remains rather unchanged regardless of the petal size.

4.1 Principal Component Analysis

To compare how well the new dimensions classify, principal component analysis are conducted. The same technique is used to generate eigenvalues, eigenvectors, covariance matrix, and coefficients to the eigenvalues.

Figure 10 shows the eigenvalues. As expected, when 2 dimensions are combined into one, the eigenvalues are relatively higher than the other 4 original attributes. Figure 11 shows the combined eigenvectors with the original attributes. The first eigenvector represents 77.6% of the variance of the data, while the first two eigenvectors represents a total of 99.1% of the variance. The first two eigenvector almost represents the whole variance of the data.

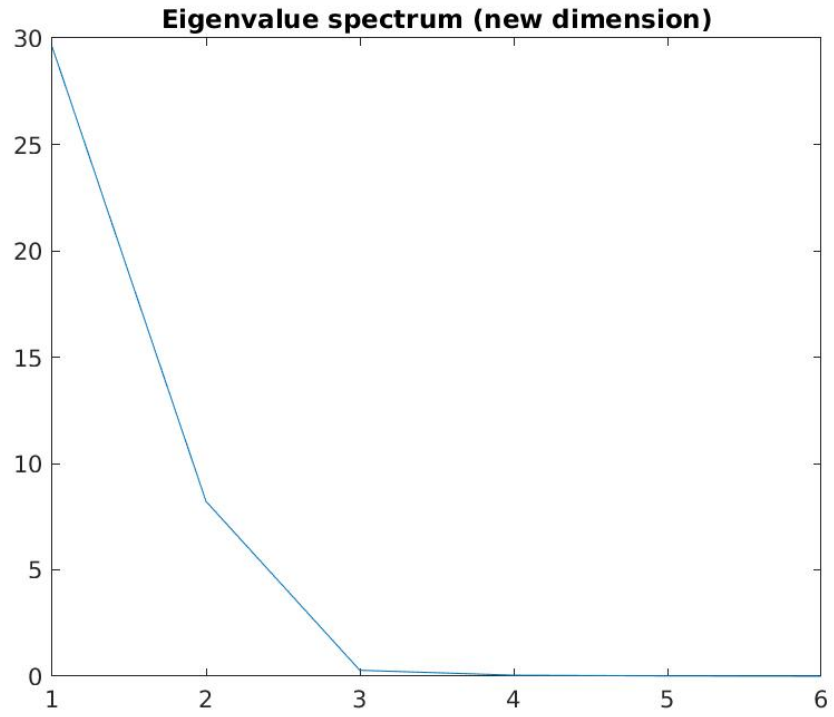

Figure 10: Eigenvalues for all attributes including the sizes of petals and sepals.

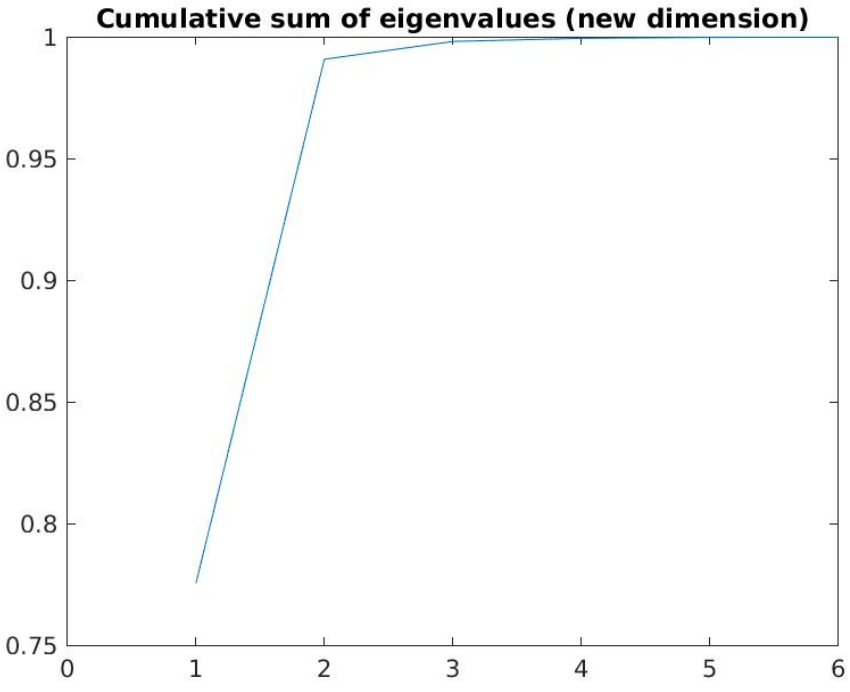

Figure 11: Cumulative sum of all eigenvalues.

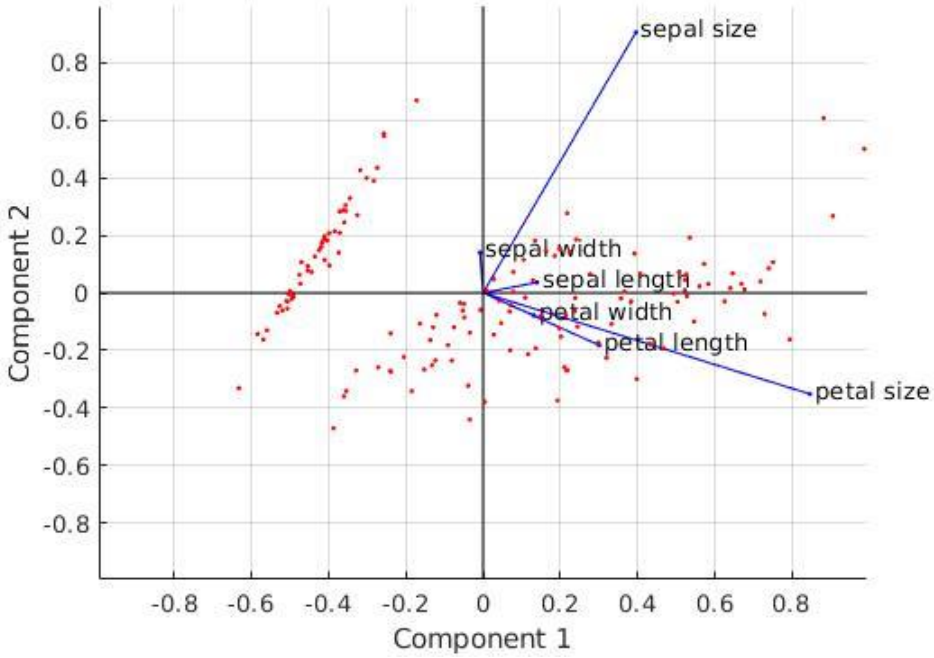

Figure 12: Biplot to visualise the coefficients of the eigenvalues and principal component scores for each data point. This biplot considers all dimensions.

Comparing the eigenvectors and eigenvalues of attributes on their own and two attributes which are derived from a combination of different attributes is unfair since each of the attributes plays a significant role in determining the class of the plant. This comparison is done was to indicate that combining attributes is possible and it might provide a clearer determinant of the class to be predicted.

5 Conclusion

The first stage of pre-processing is always crucial, to standardise the data to be in consistent format and notations. Besides, to scale and label attributes or classes to be used in principal component analysis and coordinate projections. The 2-D projections are useful when analysing data with small dimensions as the projections can shows the relationship between 2 or 3 attributes. The projections are clear and can be used to predict certain attributes based on the values of other attributes. As the dimensions get bigger, the 2-D projections can no longer be useful since it is not practical to compare every single attribute.

Hence this is where principal component analysis can be useful. It reduces the data down to its basic components while removing any unnecessary or least important parts. Eigenvectors and eigenvalues indicate the principal components of the data [5].

Another interesting finding specific to the Iris plant data set was to combine the attributes to form another attribute. The lengths and widths indicate the size of the object measured. The principal component analysis done with the new dimension found to be useful.

6 References

[1] En.wikipedia.org. (2018). Iris (plant). [online] Available at: https://en.wikipedia.org/wiki/Iris_(plant) [Accessed 28 Feb. 2018].

[2] En.wikipedia.org. (2018). Iris flower data set. [online] Available at: https://en.wikipedia.org/wiki/Iris_flower_data_set [Accessed 28 Feb. 2018].

[3] Uk.mathworks.com. (2018). Principal component analysis of raw data - MATLAB pca - MathWorks United Kingdom. [online] Available at: https://uk.mathworks.com/help/stats/pca.html [Accessed 28 Feb. 2018].

[4] Qlucore.com. (2018). The benefits of Principal Component Analysis (PCA) | Qlucore. [online] Available at: https://www.qlucore.com/news/the-benefits-of-principal-component-analysis-pca [Accessed 28 Feb. 2018].

[5] George Dallas. (2018). Principal Component Analysis 4 Dummies: Eigenvectors, Eigenvalues and Dimension Reduction. [online] Available at: https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-einenvalues-and-dimension-reduction/ [Accessed 28 Feb. 2018].

7 Appendix

The MATLAB code and processed data file can be found at: https://github.com/nzulhilmi/IDA_1446348