**Executive Summary**

Recently, three distinct sub-species of a new type of flower called the Iris have been discovered in a far-away land: Iris setosa, Iris versicolor and Iris virginica. The related information, including the sepal length, sepal width, petal length and petal width of 150 samples, 50 observations per sub-species, is wrapped in the Anderson’s Iris Dataset.

Based on the dataset, our team are interested in the key question: how similar or otherwise the three species of iris are. In order to answer to the pairwise question, our team conducted the statistical analysis in 4 steps: 1) perform an Exploratory Analysis to get a basis understanding of the data; 2) use Principal Component Analysis based on both covariance and correlation, display PC scores; 3) perform Hotelling’s T2 statistic to prove the similarity between the species; 4) perform the Multivariate Linear Model to see the effects of the features of different species in the dataset.

This report summarizes the statistical modeling and analysis results associated with the Anderson’s Iris dataset. Most of the analysis is conducted by leveraging the R studio with various statistical calculation packages. The plots and charts are from R studio directly and the tables are from Excel.

Through the statistical analysis of the Iris data, the key findings are as follows:

1) From the analysis and comparison between the covariance-based PCA and the correlation-based PCA, we recommend adapting the covariance-based PCA with the largest explanatory power and keep the first component. But we still need help from domain experts to interpret the exact conclusion.

2) From the interpretation of the Hotelling’s T-Squared statistic test, we conclude that the species have different features, which could distinguish the species from the statistically perspective.

3) From the Multivariate Linear Model, we find the coefficients and contributions of each feature to distinguish the species based on the Setosa.

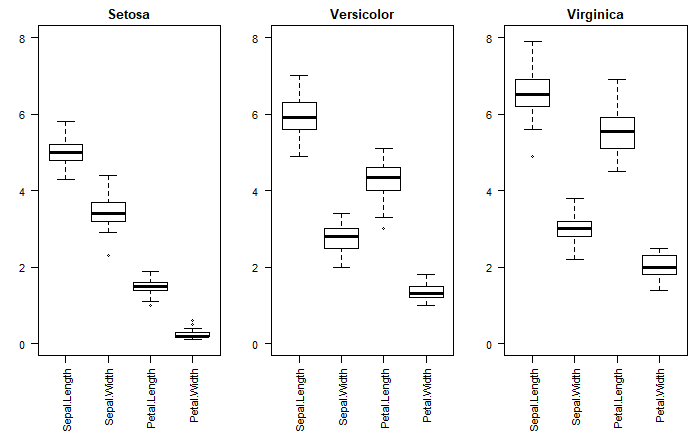
# **Exploratory Analysis**

* 1. **Basic summary of the dataset**

 In the Anderson’s Iris dataset, there are 5 variables (sepal length, sepal width, petal length, petal width and species) and 150 observations. To get a first impression of the dataset, the following figure displays 10 observations randomly selected from the Iris dataset, the first 4 variables about length and width are numeric while the “species” is a category factor with “Setosa”, “Versicolor”, and “Virginica” to represent different iris sub-species (Figure 1.1).

*Figure 1.1. A Glimpse of the Iris Dataset Variables and Observations*

* 1. **The symmetry and sparsity of the variables**

 A boxplot is a standardised way to display the distribution of a dataset based on a five-number summary: “minimum”, 1st quantile (25%), 2nd quantile (50%), 3rd quantile (75%) and “maximum” (Galarnyk, 2019). In this way, our team utilise the boxplot to display the statistical summary of the four features as the x-axis with three separate species to see how symmetric and sparse the dataset is.

*Figure 1.2. Boxplots of Iris Dataset of Each Sub-species*

From Figure 1.2, it’s clear that Setosa species has the smallest pedal size and short but wide sepal with tightly grouped data points, but the features of Versicolor and Virginica are quite similar, which require further analysis to distinguish.

* 1. **The distribution of each variables**

A histogram represents a distribution of numerical data and provides an estimate of the probability distribution. Based on the previous analysis of the boxplots, we can derive more descriptive information of each features (Figure 1.3).

|  |  |
| --- | --- |
|  |  |
| *Figure 1.3.A The Histogram of Sepal Length* | *Figure 1.3.B The Histogram of Sepal Width* |
|  |  |
| *Figure 1.3.C The Histogram of Petal Length* | *Figure 1.3.D The Histogram of Petal Width* |

*Figure 1.3. The Histograms of Features of Each Sub-species*

1. Sepal length

The sepal lengths of Setosa are very close and all the data points fall in the range from 4 to 6, while sepal lengths of Versicolor and Virginica are of wider range, from 4 to 8, and Virginica has higher probability to see longer sepal (Figure 1.3 A).

1. Sepal width

The distribution of sepal widths of Versicolor and Virginica are quite similar and all about 2 to 4, while Setosa has higher probability to see wider sepal (Figure 1.3 B).

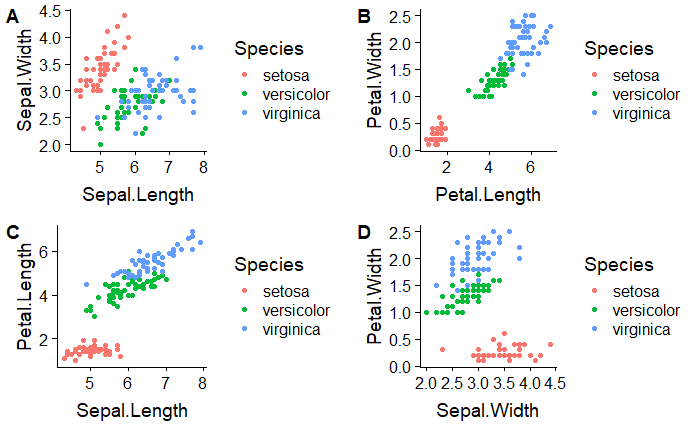
1. Petal length

The petal lengths of Setosa are very close and all the data points fall in the range from 1 to 2. Versicolor has the second longest petal at about 3 to 6, while Virginica has higher probability to see longer petal of 4 to 7 (Figure 1.3 C).

1. Petal width

The petal widths of Setosa are very close and the data points are from 1 to 2, which means the Setosa has the least wide petal. Versicolor has the second widest petal, while Virginica has the widest petal with an average of 6 (Figure 1.3 D).

* 1. **The relationships among the variables**

 Using the scatterplots, we try several types of the combination of features, to roughly estimate if there is any linear relationship or distinguishing patterns. From the Figure 1.4 (B) (C) & (D), we can find out that the Setosa has the smallest petal, which creates some spare room between the red dots and other dots, may help us to separate Setosa from other species. We will conduct further statistical analysis to see the significance, prove the relationship and find the condition to make the decision.

*Figure 1.4. The Scatterplots of Variables of Each Sub-species*

# **Principal Components Analysis**

1. **Covariance-based Principal Components Analysis**
   1. Methodology

In this case, from the previous exploratory analysis, we get some descriptive information of the features of different iris species. It seems that petal length and petal width are most distinct for the three species and could help us to distinguish the species.

The second step of this statistical analysis is to perform Principal Components Analysis (PCA), which would allow us to systematically discover the patterns between features (variables) and the species. A key objective of the Principal Component Analysis is to find the linear combination of the original variables that maximises the variability.

The basic steps to perform a covariance-based Principal Component Analysis in R studio are as follows: a. prepare a data frame that holds the numerical columns of interest (sepal length, sepal width, pedal length and pedal with as the response vectors) and set the species column aside; b. scale the data to 0 mean, calculate the unit variance, and come up with the covariance matrix; c. perform the PCA with the function prcomp() in R studio and display results.

* 1. Results

The main results from Principal Component Analysis in R studio are standard deviations and a rotation matrix with PC scores. The interpretation of the Principal Component Analysis is very speculative and can’t derive a specific conclusion. We need to get more data and discuss the interpretation with domain experts.

1. The Eigenvalues and PC Scores.

 From the eigenvalues table, the first principal component has the eigenvalue larger than 1 and could explain 92.46% of variation in the data (Figure 2.1).

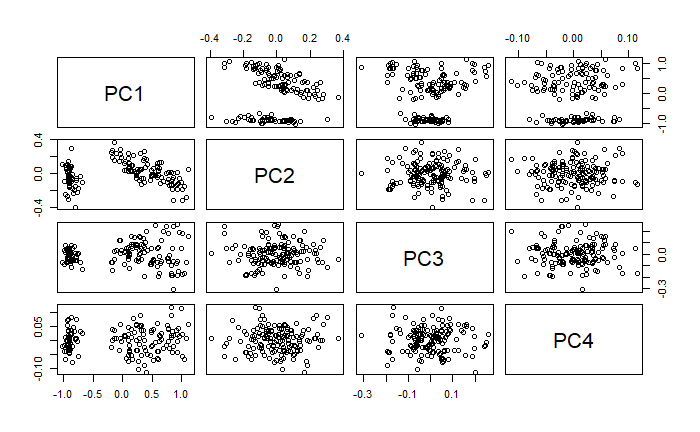
*Figure 2.1. The Eigenvalues of Covariance-based Principal Components Analysis*

 To summarise, the PC Scores are re-arrangements of the data in a form that allows to explain the data set with less variables, which refers to the dimension reduction in the statistics. The PC scores represent how much each variable relates to the component.

*Figure 2.2. The Rotation Matrix of Covariance-based Principal Components Analysis*

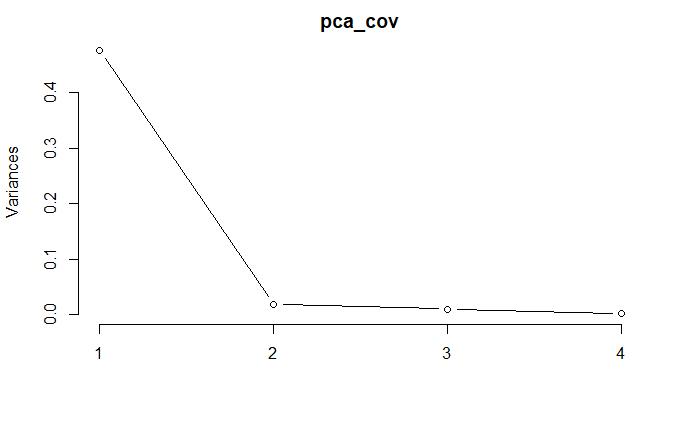
From the rotation matrix, we can see that the first principal component has large positive associations with petal length and petal width, while the second principal component has large negative association with sepal width (Figure 2.2). We will conduct the further interpretation and analysis based on the PC scores.

1. The Scatterplots of PC Scores.

 From the Scatterplots, we can roughly interpret some basic conclusions: The plots don’t appear as random scatter. From PC1 plots on the left and on the top (they are symmetric and display the same patterns), it’s clear that some of the dots are separated from the others (Figure 2.3).

*Figure 2.3. Scatterplot of PC Scores of Covariance-based Principal Components Analysis*

1. The scree plot of the components.

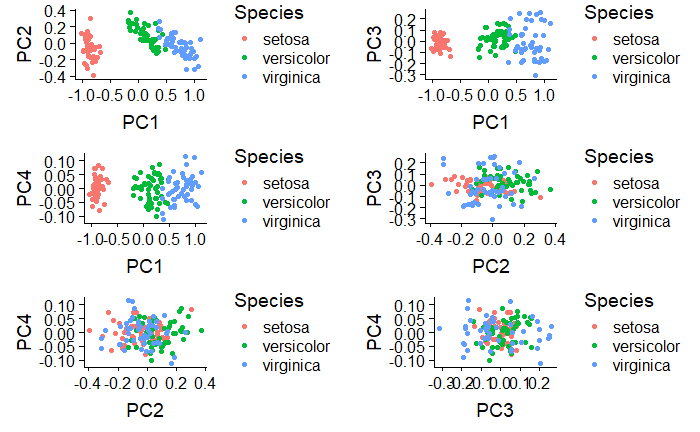
From the Scree Plot of the components, the first component could explain more than 90% of the variances in the data, which is the only component we’d like to keep.

*Figure 2.4. Scree plot of the Components of Covariance-based Principal Components Analysis*

1. PC scores comparison

Followings are the scatterplots of six pairs of PC scores with species as different colors, which would help us to find out the patterns (Figure 2.5).

The first principal component has significant negative associations with the Setosa species, 0 to 0.5 associations with the Versicolor species, and significant positive associations with the Virginica species.

As for the rest three components, the associations are not significant. In this way, we will keep the first principal component from the covariance-based Principal Component Analysis.

*Figure 2.5. Scatterplots of the PC Scores of Covariance-based Principal Components Analysis*

1. Correlation-based Principal Components Analysis
2. Methodology

The second method of Principal Components Analysis is the correlation-based PCA. The basic steps to perform a correlation-based Principal Component Analysis in R studio are quite similar to the steps of covariance-based PCA with some adjustments on the parameters only.

1. Results
2. The Eigenvalues and PC Scores.

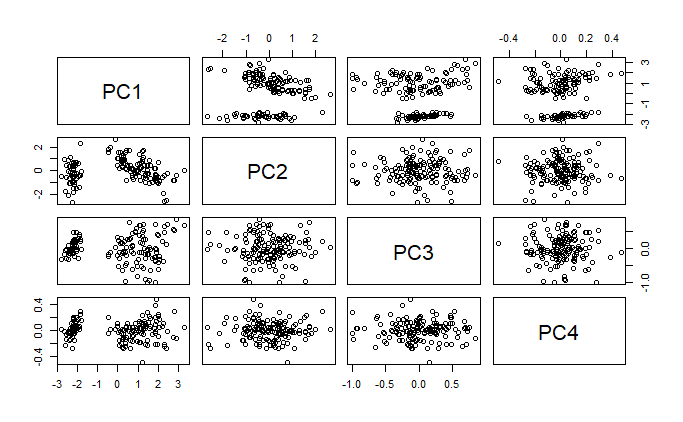
*Figure 2.6. The Eigenvalues of Correlation-based Principal Components Analysis*

From the eigenvalues table, the first principal component has the eigenvalue larger than 1 and the eigenvalue of the second component is almost 1. The first and the second components could explain 83.81% of variation in the data (Figure 2.6).

 From the rotation matrix, we can see that the first principal component has large positive associations with sepal length, petal length and petal width, while the second principal component has large negative association with sepal width (Figure 2.7). We will conduct the further interpretation and analysis based on the PC scores.

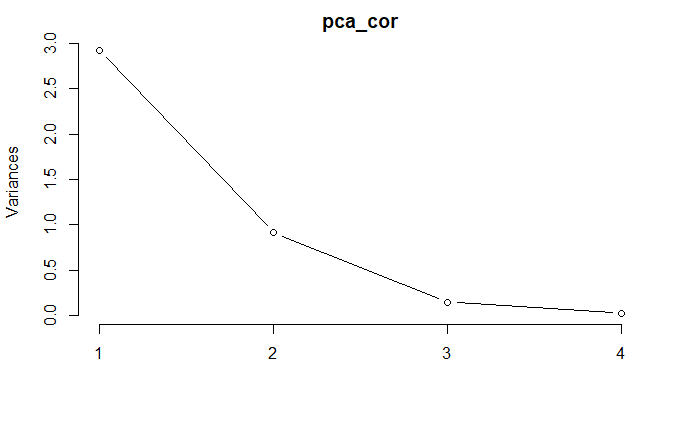
*Figure 2.7. The Rotation Matrix of Correlation-based Principal Components Analysis*

1. The Scatterplots of PC Scores.

From the Scatterplots, the basic conclusion is the same as before: The plots don’t appear as random scatter. From PC1 plots on the left and on the top, it’s clear that some of the dots are separated from the others (Figure 2.8).

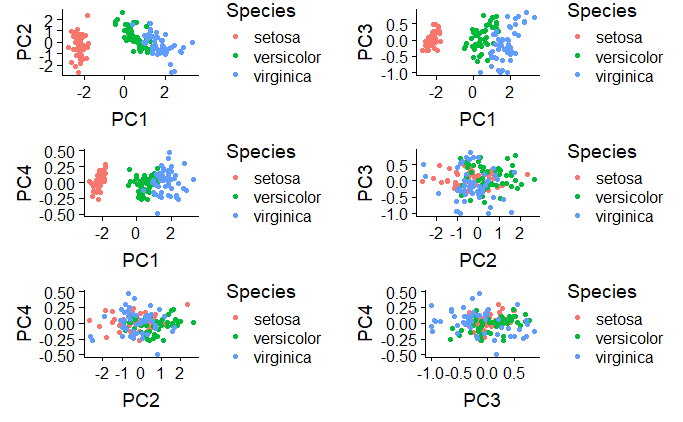
*Figure 2.8. Scatterplot of PC Scores of Correlation-based Principal Components Analysis*

1. The scree plot of the components.

From the Scree Plot of the components, the line of first component and the second are steep, which means these two components are sensitive. Different from previous conclusion, we’d like to keep both of the components.

*Figure 2.9. Scree plot of the Components of Correlation-based Principal Components Analysis*

1. PC scores comparison

Followings are the scatterplots of six pairs of PC scores based on correlation PCA with species as different colors, which would help us to find out the patterns.

*Figure 2.10. Scatterplots of the PC Scores of Covariance-based Principal Components Analysis*

The first principal component has significant negative associations with the Setosa species, but the boundary between Versicolor species and Virginica species is not clear.

As for the rest three components, the associations are not significant. In this way, we will keep the first principal component from the covariance-based Principal Component Analysis.

1. **Comparison of Covariance-based and Correlation-based PCA**

The similarities between the covariance-based PCA and the correlation-based PCA:

1. The first principal component has significant association with petal length and petal width.
2. The eigenvalue of the first principal component is larger than 1 and can explain more than 80% of variance of the data.
3. The second principal component has large negative association with sepal width.

The differences between the covariance-based PCA and the correlation-based PCA:

1. In the covariance-based PCA, the first principal component has larger eigenvalue and better explain of the variance in the data.
2. In the correlation-based PCA, the second principal component has stronger explanatory power, but the total performance is worse than the covariance-based PCA.

According to the performance of both models, the covariance-based PCA can provide more information with stronger explanatory power with 92.46% from the first principal component. But as mentioned before, the interpretation of the PCA results is speculative, and requires domain experts help to come up with the final conclusion.

# **Species Pairwise**

1. Methodology

To compare the species pairwise, we adapt the Hotelling T-Sqaured statistics, which is the multivariate counterpart of the T-test. For the Iris dataset, we have four variables to describe the iris species, which requires Hotelling’s T-Squared to test the similarity of the samples.

In this case, we use sepal length, sepal width, petal length and petal width as the variables to explain the iris species and we have three pair of two-sample test: Setosa vs. Versicolor, Setosa vs. Virginica, and Versicolor vs. Virginica.

The Hotelling’s T-Squared has the following advantages: 1) the Type 1 error is controlled, 2) the relationship between multiple variables is taken into consideration, and 3) can generate an overall conclusion and summarise the between-group differences (Statistics How To, 2019).

1. Results

Leveraging the package in R studio, we can easily get the result of the Hotelling’s T-Squared result of Setosa vs. Versicolor is 625.46, Setosa vs. Virginica is 1182.6, and Virginica vs. Versicolor is 86.148, which are much higher the target F-statistic, with 99.9% confidence level. The conclusion is to reject the null hypothesis that the variables are not significantly different, which means the three iris species are different, from the statistical aspect with 150 sample tests.

# **Multivariate Linear Regression**

1. Methodology

The Multivariate Linear Regression method uses sepal length, sepal width, petal length and petal width to explain the species of Setosa, Versicolor and Virginica. The method is to find the coefficients of each variables to minimise the total error and the coefficients could be considered as the attribution of each variables.

1. Results

R studio helps us to calculate the coefficients of each variables towards different species based on Setosa (Figure 4.1).

The numbers in the table tell us the following information:

1. Setosa has average sepal length of 5.0060, while Versicolor has longer sepal by 0.93 and Virginica has longer sepal by 1.5820.
2. Setosa has average sepal width of 3.4280, while Versicolor has smaller sepal by 0.6580 and Virginica has smaller sepal by 0.4540.
3. Setosa has average petal length of 1.4620, while Versicolor has longer sepal by 2.7980 and Virginica has longer pedal by 4.090.
4. Setosa has average petal width of 0.2460, while Versicolor has larger petal by 1.0800 and Virginica has larger petal by 1.7800.

As there’s still the unavoidable statistical error, the linear model represents the ideal situation and the true situation may be slightly different.

*Figure 4.1. Scatterplots of the PC Scores of Covariance-based Principal Components Analysis*



# **Reference**

[1] Galarnyk, M. (2019). *Understanding Boxplots*. [online] Towards Data Science. Available at: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 [Accessed 3 May 2019].

[2] Statistics How To. (2019). Hotelling's T-Squared: Simple Definition - Statistics How To. [online] Available at: https://www.statisticshowto.datasciencecentral.com/hotellings-t-squared/ [Accessed 5 May 2019].