# ENGSCI 331

## Eigenproblems Assignment Report

### Author: Ghassan Al-A’Raj

### ID: 434557967

Table of Contents

[1. Introduction 2](#_Toc48676652)

[2. Implementation and Testing 2](#_Toc48676653)

[2.1. High-Level implementation 2](#_Toc48676654)

[2.2. C++ implementation 3](#_Toc48676655)

[3. Principle Component Analysis 4](#_Toc48676656)

[3.1. Explained Variance Ratio Analysis 4](#_Toc48676657)

[3.2. Pair of PCs 5](#_Toc48676658)

[3.3. Signal reconstruction 7](#_Toc48676659)

# Introduction

We are working for Auckland based technology start-up Coffee-AI, which develops an intelligent coffee machine named e-Barista. This coffee machine is supposed to automatically tune its grinding and brewing process according to the coffee beans provided. The heart of each e-Barista is a Fourier-Transform Infrared Spectrometer, which measures the infrared spectrum of absorption of the coffee beans. We are asked to compute the eigenvectors of the cross-correlation matrix of the reflectance spectra to try and find the a pair of eigenvectors that best separate the two types of coffee beans we are trying to analyse, Arabica coffee beans and Robusta coffee beans.

# Implementation and Testing

## High-Level implementation

To better understand the algorithms, we had to implement. We firstly coded the power method and deflate functions in python, we then tested the functions with a variety of test cases. This step was crucial to translating the algorithms from a set of mathematical expressions to set of programmatical instructions. Thus, the python implementation was used as a blueprint for the C++ implementation.

Note that for the python implementation we used a test-driven design philosophy i.e. the test cases were written first, then the doc-strings for each function, then finally the functions. I believe this methodology of implementation allowed us to gain the most comprehensive understanding of the algorithms. Thus, making the C++ implementation that much easer to code and test.

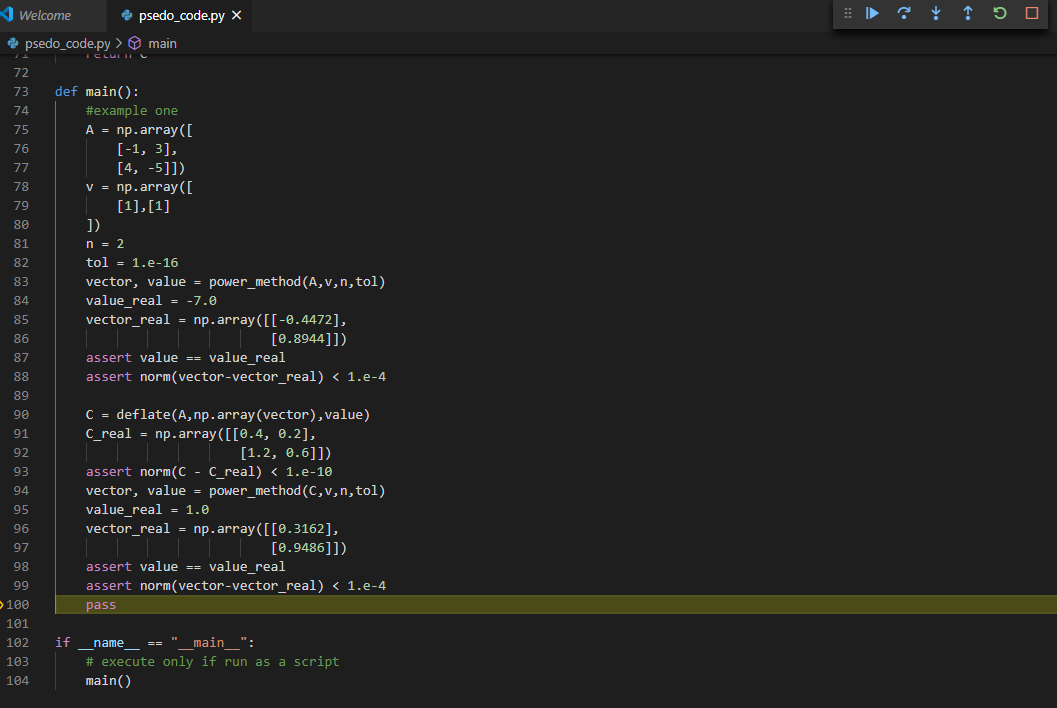


Figure 1 psedo\_code.py passing test cases

## C++ implementation

Implementing the power method and deflate function were easy enough, because of how useful the python implementation turned out to be. The implementation of the other two functions (Centre matrix and Covariance matrix) proved more difficult as we didn’t implement them in a higher level language before we started coding them in C++, but the same test-driven design philosophy was used and the overall code is efficient and functional. Using the functions to get the eigenpairs and saving to a csv file was very simple and did not require much in the way of testing or time.

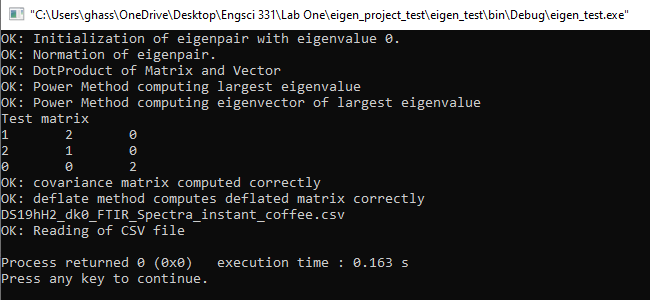


Figure 2 eigen\_functions passing test cases

# Principle Component Analysis

## Explained Variance Ratio Analysis

After extracting the eigenpairs from the centred covariance matrix using the C++ code and exporting the results in the form of two csv files aptly labelled ‘eigenValues.csv’ and ‘eigenVectors.csv’ respectively. we then imported the data into MATLAB, we can conduct an explained variance ratio analysis, the results of which can be seen below.

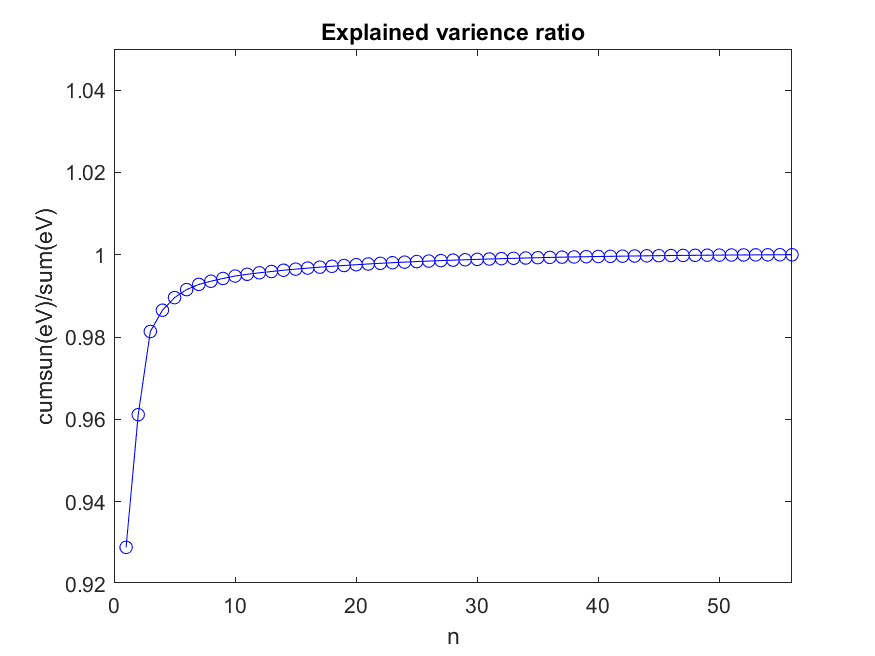


Figure 3 explained variance ratio graph

Following the advice given in lectures, I looked for the step in which the most change occurred. I determined that to be 3rd step i.e. the first 3 principle components were to be used to reduce the dimensionality of the data. After projecting the data onto the principle components. The points were separated into two groups, the first 29 belonging to Arabica coffee beans and the last 27 belonging to Robusta coffee beans.

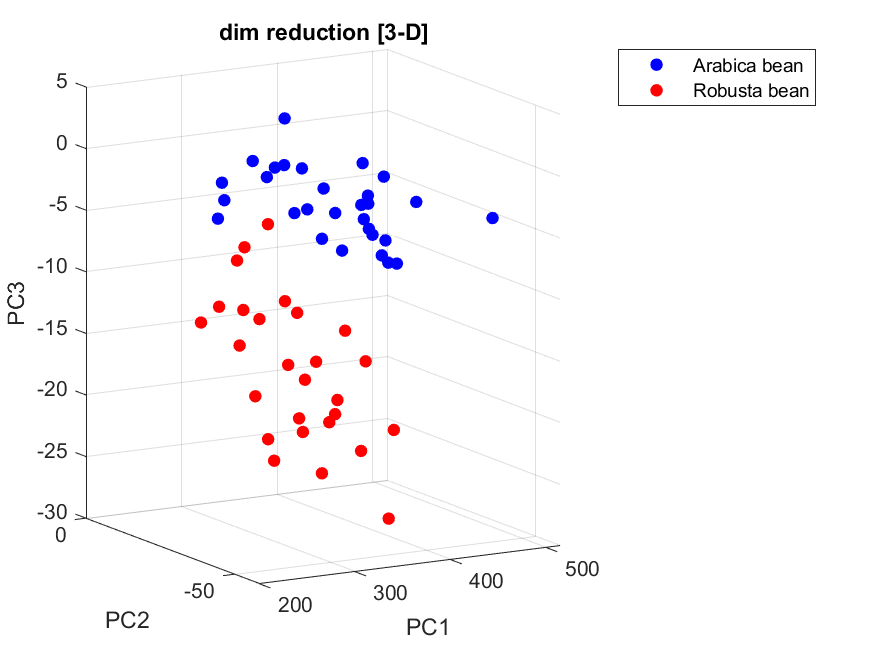


Figure 4 3-D scatter plot showcases the 56 spectra projected onto the first 3 PCs

As we can see from the above figure, projecting the data onto the first 3 principle components provides a good separation between the two types of coffee beans.

## Pair of PCs

To identify the pair of principle components which provide a good separation for the two types of coffee beans I implemented a ‘soft’ metaheuristic to find to find the or 50 so combinations that would achieve reasonable results. Firstly I found all set of combinations possible (1540 combinations in total), I then found the projection of the spectra onto that set of principle components, then I separated the values into the two types of coffee beans and then compared the distance of each Arabica point to each Robusta point, if the distance was greater than some arbitrary value, that observation was set as true (it has a value of 1). Summing all true observations, gave us a measure of the fitness of each combination. I plotted a histogram of the scores.

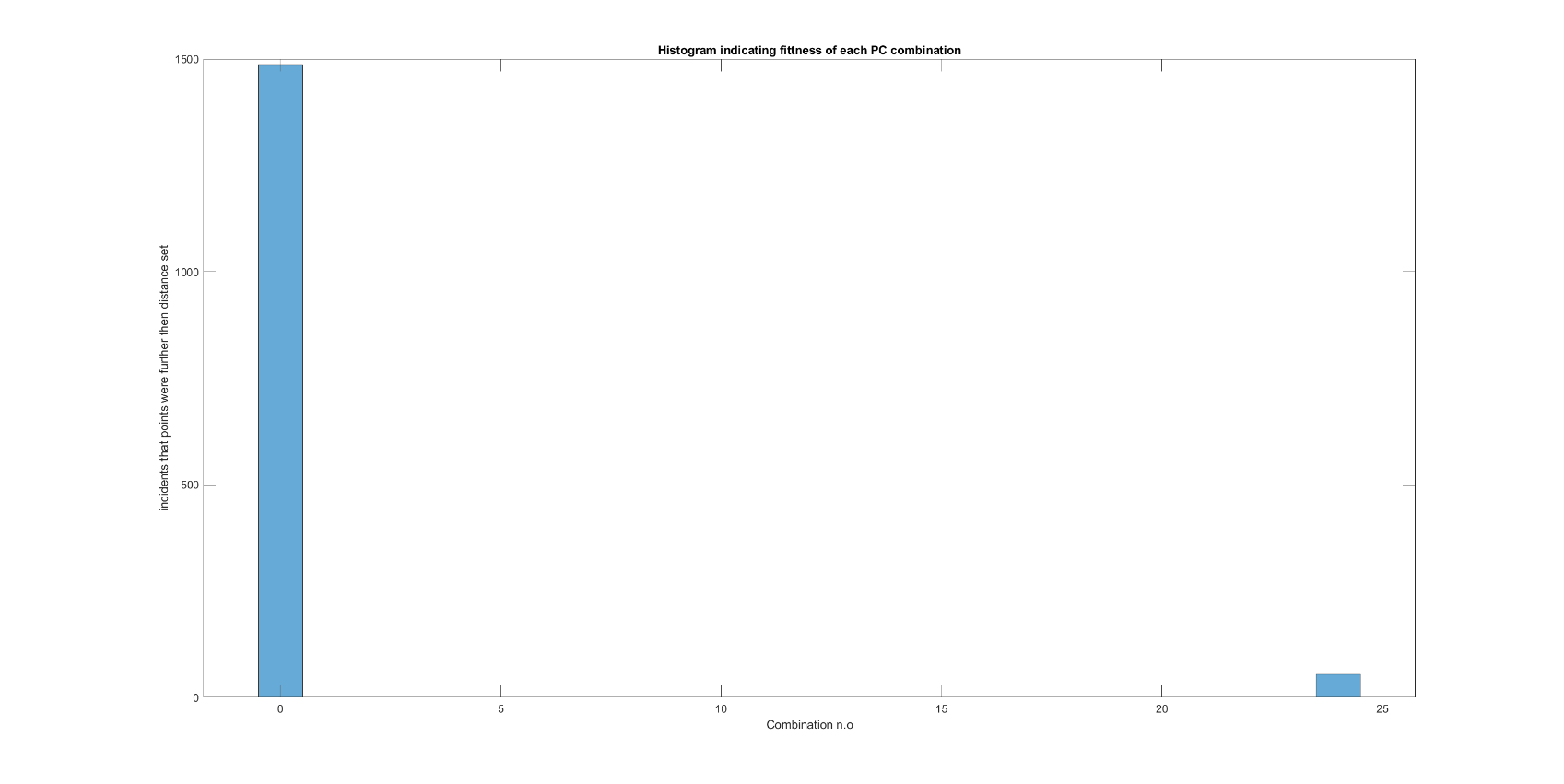


Figure histogram of the scores of each combination

While the figure above does not look impressive, it shows that we need to consider the first 10-ish combinations, interestingly we should also consider some mid-20 combinations as they have comparatively high scores. The most important take away from this graph is we should not consider combinations beyond the mid-20’s, as they failed to even score even 1 point.

To ensure my results make sense I plotted a scattering of the projection points generated from the 823rd combination, which contained the 18th and 51st principle components. This combination had a very low distance in both the x and y directions, at 6.4021 and 1.7368 units in each respect direction. We expect to see a very bad separation between the two types of coffee beans and as seen in the graph below, there is no clear separation and thus this combination and similar combinations are discarded.

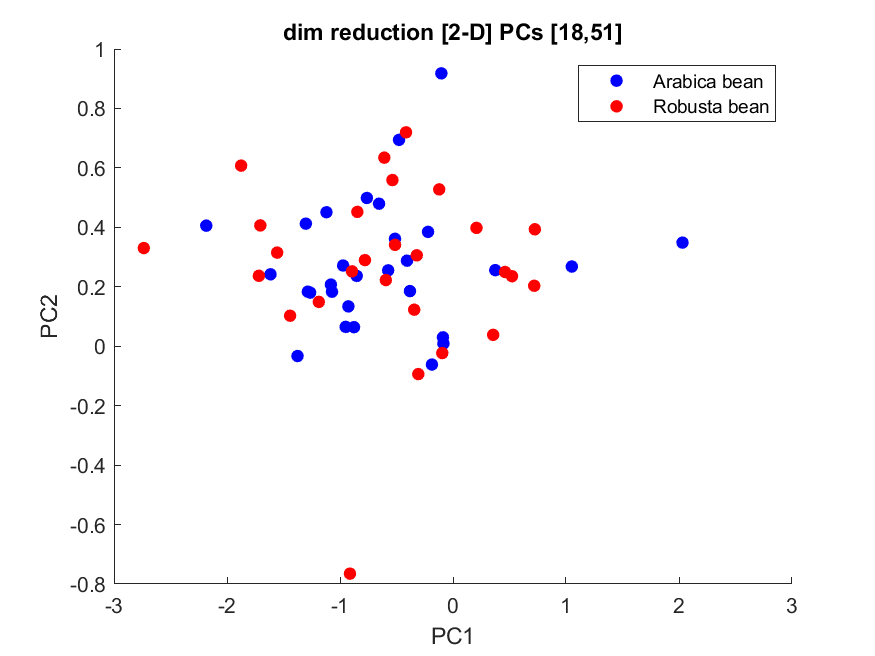


Figure 6 2-D scatter plot showcases the 56 spectra projected onto the 18th and 51st PCs

After discarding the ‘bad’ combinations we reduce the number of combinations from 1540 combinations to 10-ish reasonable ones, coincidentally this is the first 10 combinations. This is expected as the power method finds the most dominant eigenpairs first, but all 10 have very similar scores. Inspecting them we find that projecting on the first and third principle components offers the most separation, looking at the graph below we see that we can clearly separate between the two types of coffee beans.

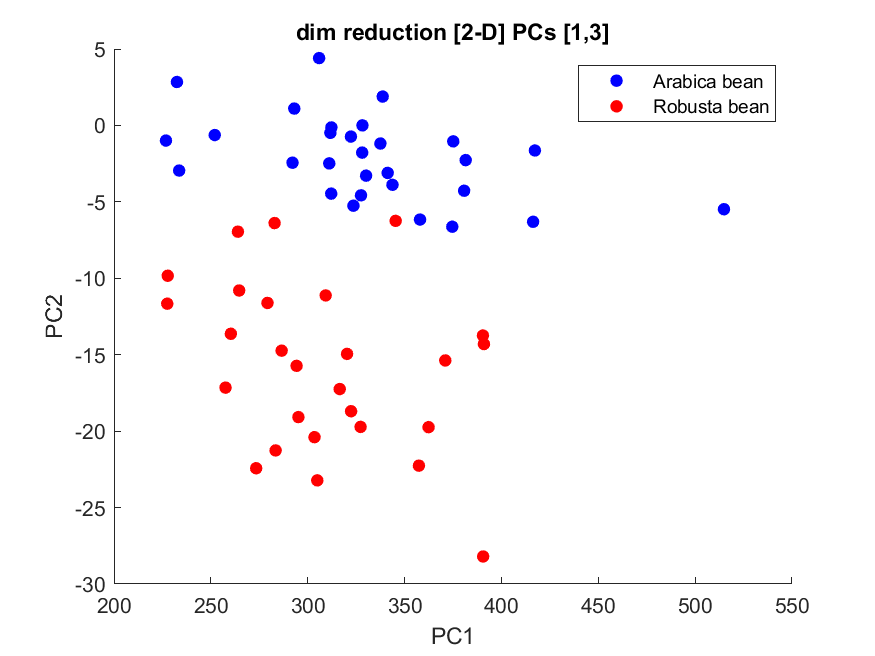


Figure 7 2-D scatter plot showcases the 56 spectra projected onto the 1st and 3rd PCs

## Signal reconstruction

We used the 6 most important principle components i.e. the first 6 principle components due to the power method. We then used the eigenvectors and the points corresponding to the first spectra of Arabica beans and first spectra of Robusta beans to reconstruct said spectra the results of which are below.

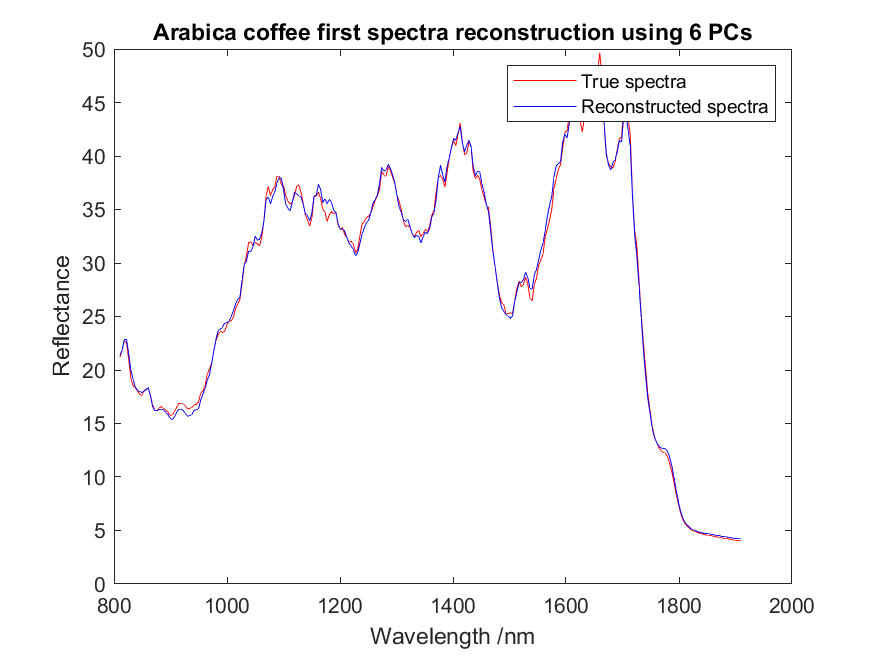


Figure 8 Arabica coffee signal reconstruction

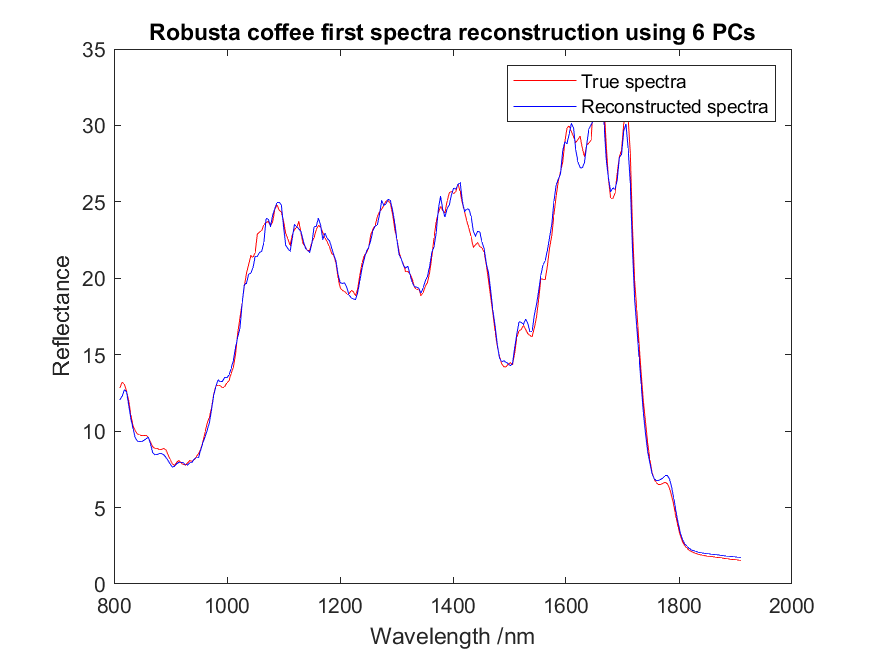


Figure 9 Robusta coffee signal reconstruction