Introduction to Statistical Learning Coursework

Tudor Trita Trita
CID 01199397

February 18, 2020

This is my own work unless stated otherwise. The plots and code used to create them are adaptations of the code given in the lectures.

1 Report

1.1 Introduction

In this report, we explore the use of distance-based methods for the analysis of data from the 27 European Union (EU) member states taken between 2018 and 2020. This is done from the viewpoint of a data analyst working for a company specialising in giving consulting advice to organisations that have been negatively affected by the UK's departure from the EU and wish to explore the similarities of the economies of the remaining EU member states for potential investment or relocation after transition deal has ended.

We have compiled a dataset that consists of 38 non-geographical variables for each EU country. The data includes economic indicators such as GDP, population and income as well as other indicators such as police trust ratings, health statistics and satisfaction with the environment. The data has been gathered from the EU Open Data Portal. Special care has been taken to avoid including data which is inherently dependent on location, for example, climate or weather data.

1.2 Analysis and Methodology

We now begin the analysis. We can model our dataset as an n x p matrix X, where $n = 27, p = 38$. It is important to scale the data as we would like to give each variable equal importance. Our goal is to define some distance/dissimilarity metrics and construct an n x n distance matrix E for each of those metrics. For this analysis, we will be working with the Euclidean and Manhattan metrics. We then wish to recover a configuration Y modelled as a n x q matrix - a set of coordinates in the space of q dimensions. For our purposes, we will be working with $q = 2$ to aid with the visualisation as we can easily visualise two dimensions on a screen.

The first analysis that we will be performing is multi-dimensional scaling (MDS). This involves using Euclidean distances as our metric, giving us an Euclidean distance matrix. Figure 1 shows the eigenvalues for the multidimensional scaling, showing that there are two larger eigenvalue followed by smaller eigenvalues forming an elbow. It is also worth noting there are no negative eigenvalues, which is to be expected, as having negative eigenvalues would indicate non-Euclidean error; something that is not possible when using Euclidean distances.

We are now in a position where we can recover the configurations. Using the R command `cmdscale` and plotting the recovered configurations, we can see the results in Figure 2. It is not surprising that the configuration does not resemble a map of Europe as we never used any geographical data. Note that this is still the case even if we translate or rotate the configuration. However, we do see some resemblance between geographical proximity and closeness in this configuration e.g. the countries of Sweden, Finland and Norway. Later on, we will perform K-Means clustering to mathematically identify clusters followed by some qualitative work on what these clusters may mean.

We will now use a different distance metric and perform a similar analysis, namely Ordinal Scaling. We recover a different distance matrix and plot the eigenvalues seen in Figure 3. The plot does include some negative eigenvalues, this is due to the non-euclidean distances caused by the Manhattan metric. Figure 4 illustrates the magnitude of the positive and negative eigenvalues. Using the ordinal scaling function isoMDS in R, we plot the recovered configuration in Figure 5.

As we can see, the configurations look different at first. This is because they need to be oriented and scaled in the optimal way such that they are as 'close' as possible to each other. To achieve this quantitatively, we perform Procrustes Analysis to the configurations, having the Manhattan distance configuration as the $\hat{Y}$. The configurations are plotted superimposed in Figure 6, with the configuration using MDS is in black and the Ordinal Scaling configuration is in blue. As we can see, there are some differences but overall both configurations are relatively close to each other.

Moving on to clustering methods, we will now use K-Means to find an optimal number of clusters and centroid locations for our MDS configuration. Using the Silhouette and Elbow methods (Figures 7, 8), we see that the average silhouette width is highest at a number of clusters $k = 3$ and an elbow appears to emerge for the within-sum-of-squares plot $k = 3$. This gives evidence to choose $k = 3$ as the number of clusters.

1.3 Discussion of Results

The 'socioeconomic' distances found by projecting the data onto two dimensions is surprisingly similar to the distances found in real life. This is highlighted by the fact that with three clusterings, the countries have been clustered into three separate groups (Figure 9), which upon inspection can be broadly defined as the following: 1. (Black) countries in western Europe (France, Spain, Germany, etc.), 2. (Green) countries in eastern Europe (Czechia, Poland, Hungary, etc.) and 3. (Black) Scandinavian countries (Sweden, Finland, Norway, etc.). There are exceptions, e.g. Romania in group 1 or Austria in Group 3. These excep-

tions may arise from two ways. Firstly and most likely, our clustering method may have assigned these countries to the wrong cluster. Secondly, it could be the case that these individual countries may have unique socioeconomic circumstances that are uncommon for their location within Europe.

From a historical standpoint, the difference between groups 1 and 2 can be partly explained by the partition of Europe during the cold war into the western capitalist block and the eastern communist block. It is worth pointing out the fact that even after 30 years after the fall of the Berlin Wall, we are still able to find empirical evidence of the effects of the cold war divide on Europe. Group 3 forms it's own group likely due to cultural and economic benefits of the Nordic countries.

From a commercial point of view, this analysis serves as a guide to the similarities of the countries in the European Union. For example, an investor may wish to diversify their investments into fixed income from the different countries. For example, to gain exposure to the different economy types in the European Union, one might want to buy bonds from the central bank of 1 or more countries from each of the clusters identified above.

2 Figures

Figure 01: Eigenvalues for Multi-Dimensional Scaling

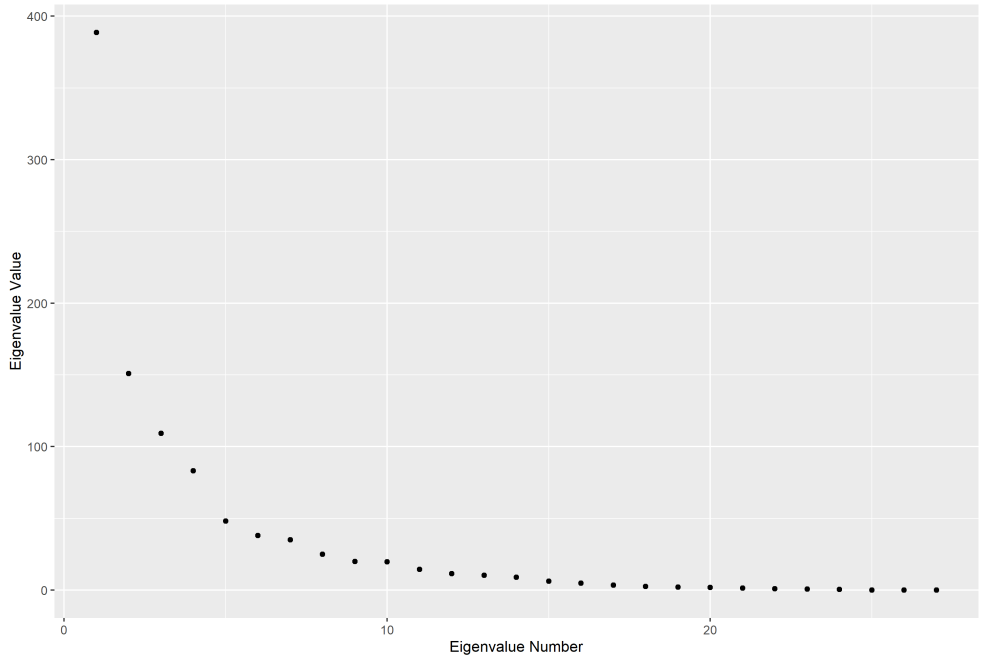

Figure 02: Recovered Configuration Multi-Dimensional Scaling

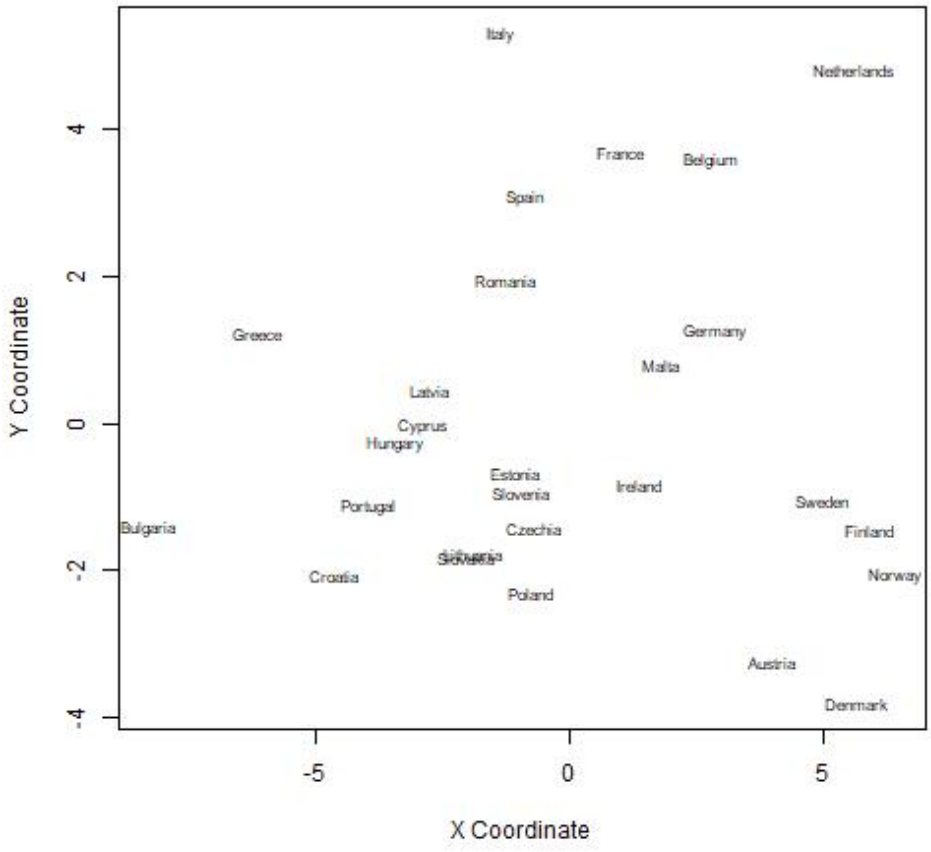

Figure 03: Eigenvalues for MultiDimensional Scaling Manhattan

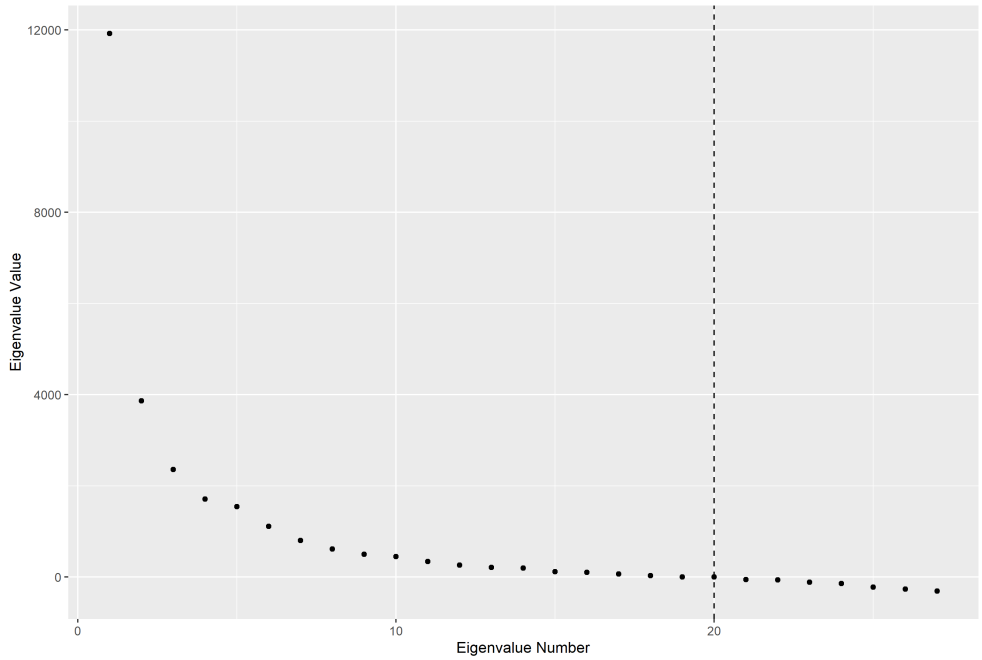

Figure 04: log(|Eigenvalues|) for Manhattan Distance

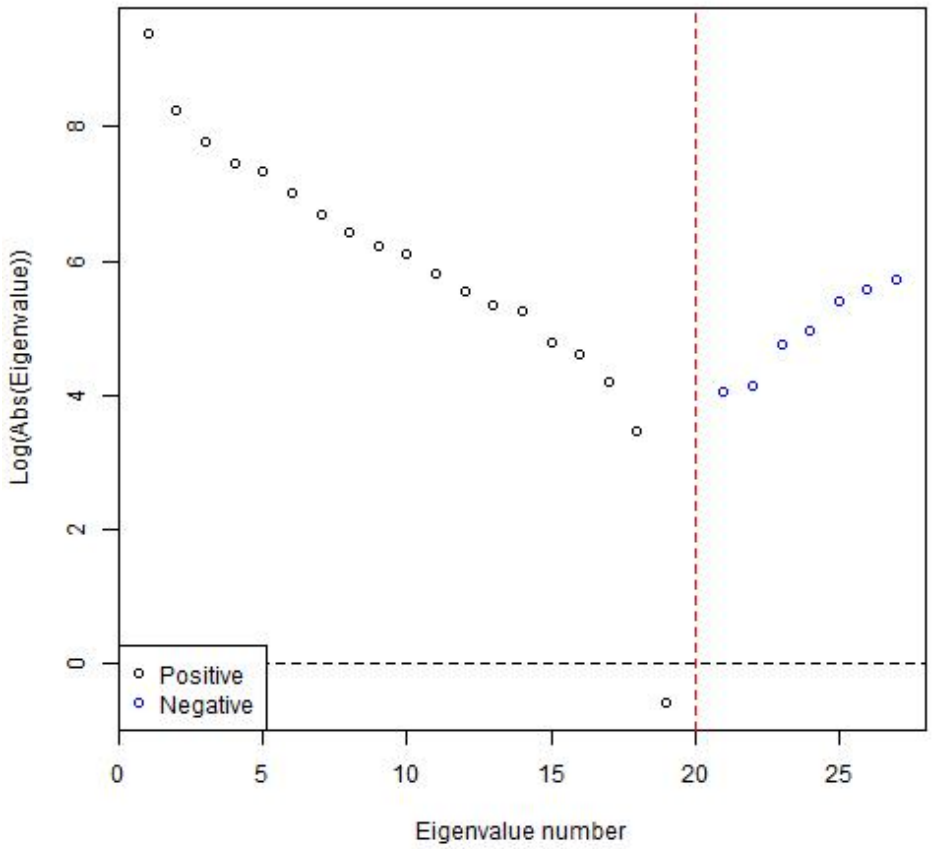

Figure 05: Configuration using Manhattan Distances

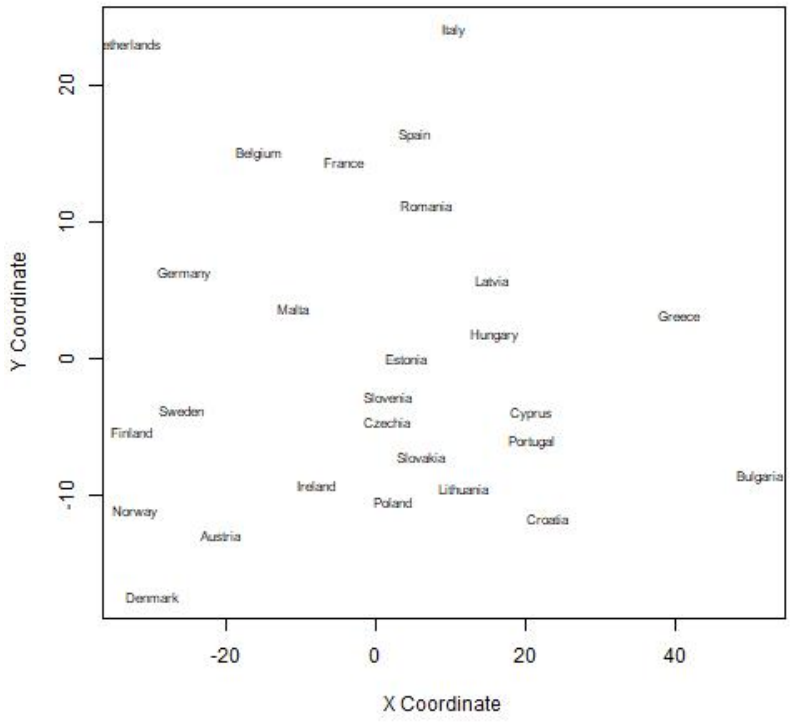

Figure 06: Plot of Both Configurations Superimposed

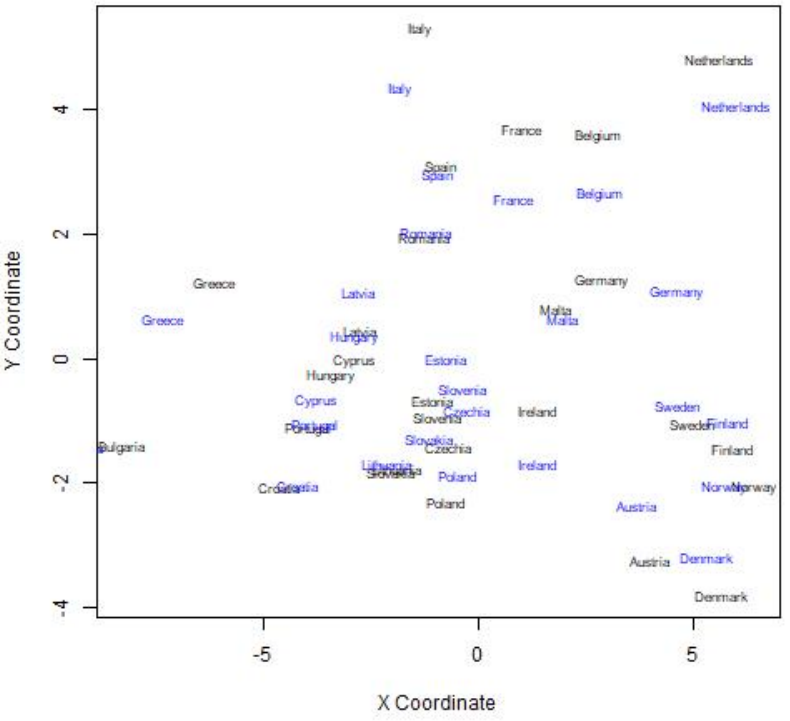

Figure 07: Silhouette Method for Clustering

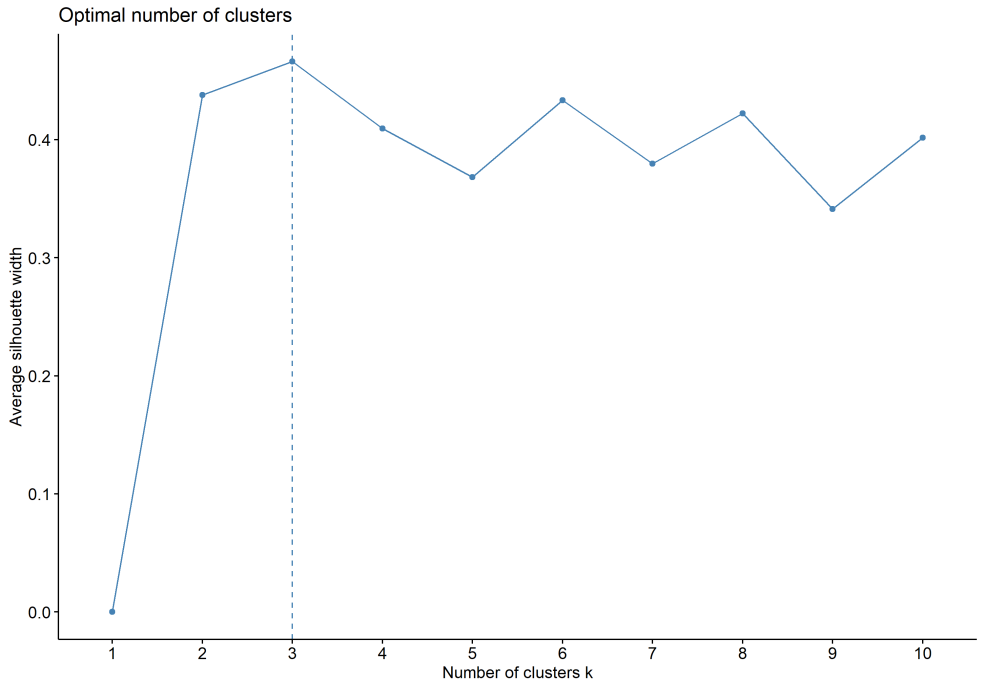

Figure 08: Elbow Method (WSS) for Clustering

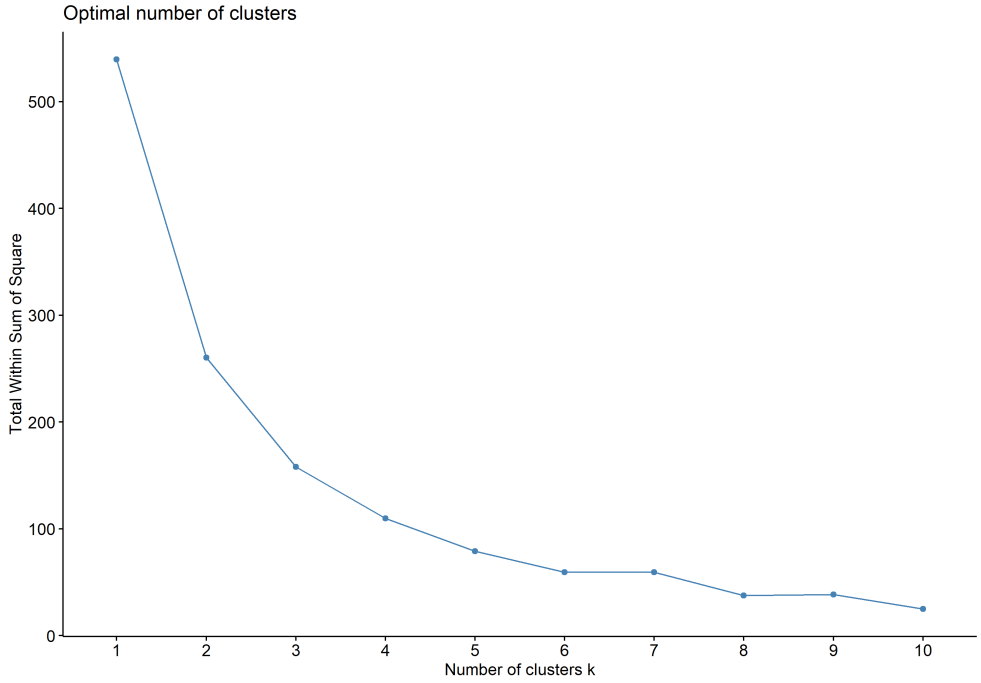

Figure 09: K-Means Clustering K=3