Multidimensional Scaling and Canonical Correlation Analysis

Gregorio Saporito

Part a: Multidimensional Scaling

(i) Implementing multidimensional scaling analysis on economic inequality among EU member states.

Multidimensional scaling was performed on economic inequality indexes retrieved from Eurostat which refer to year 2018. In particular these indexes are:

- SDG_05_30: Gender employment gap (percentage of total population)
- ILC_DI12: Gini coefficient of equivalised disposable income
- SDG_01_40: People living in households with very low work intensity (percentage of total population aged less than 60)
- TESSI082: Material Deprivation (percentage)
- TESSI180: Income quintile share ratio
- ILC_LI11: Relative at risk of poverty gap by poverty threshold

(ii) Commenting on the results, methodology and package functions used with the help of figures/tables

Just for comparison PCA was first performed on the dataset. K-means clustering was performed over the PCA output to identify similar groups. The number of clusters chosen is 4.

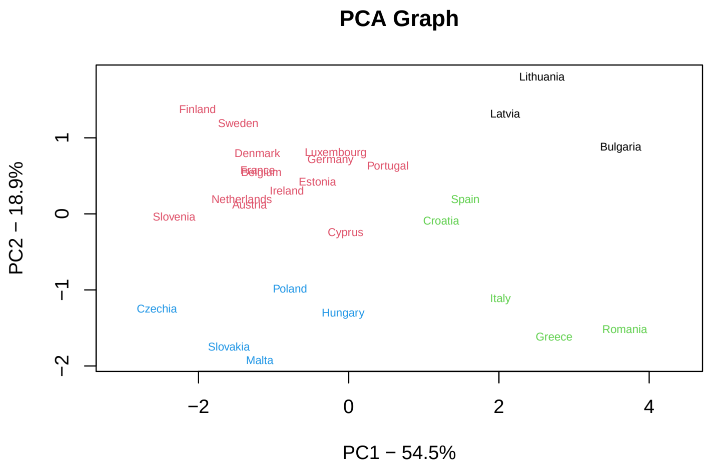

Figure 1: PCA Graph

As can be seen from Figure 1, Italy falls in the same category as Romania, Greece, Croatia, and Spain, while the red cluster includes, among others, Scandinavian countries which are generally know for having higher economic equality.

The same analysis was then performed with Multidimensional Scaling and as a distance metric the Euclidean distance. The function used for this purpose is `cmdscale()` which performs classical multidimensional scaling over a data matrix containing a distance structure. Classical MDS uses a fixed distance metric imposing a linear relationship between the samples. With Euclidean distances the output turns out to be the same as PCA.

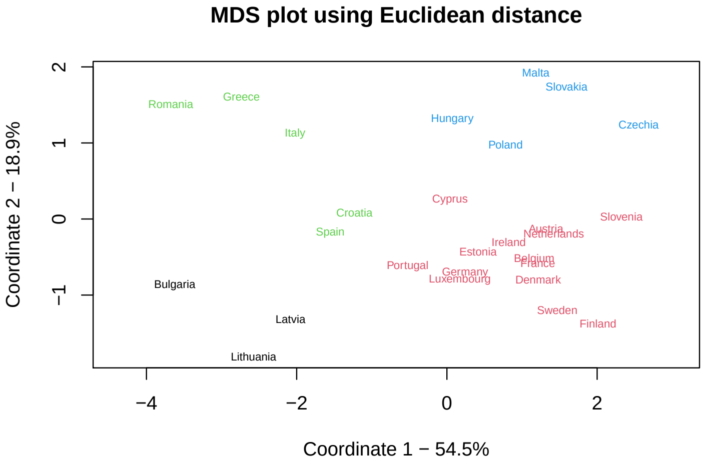

Figure 2: MDS plot

As expected, the results from Figure 2 are the same as those from Figure 1 (the results from MDS just happen to have an inverted orientation). For this reason, also the results from the cluster analysis are the same. With 2 dimensions we are able to explain 73.4% of the variabilty in the data which is a fairly satisfactory result.

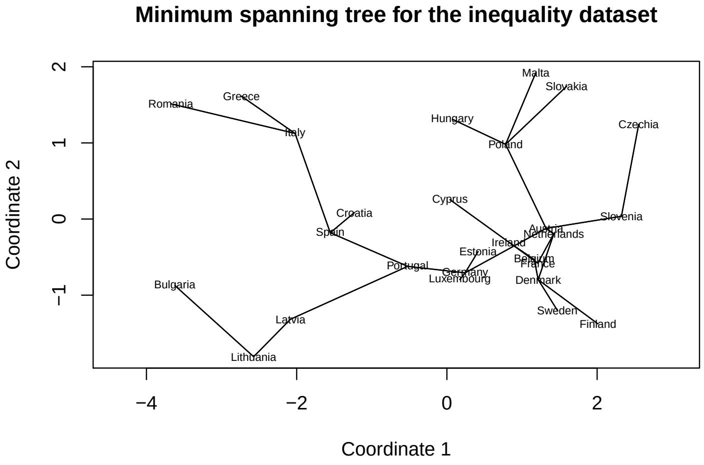

Figure 3: Minimum spanning tree

Another useful tool to visualise the relationship between the EU countries in 2 dimensions is the use of a minimum spanning tree (Figure 3) which links all the vertices without cycles with the minimum edge weight. The task is performed over the same reduced 2-dimensional space as in Figure 2. Figure 3 shows that the closest “neighbours” of Italy are Greece, Romania and Spain.

What if the underlying construct of the data is non-linear? In such case classical MDS would be less effective. In contrast, non-metric MDS allows for non-linearities in the data by finding an optimal scaling through an iterative approach. For this purpose the function `isoMDS()` was used. The same K-means procedure is applied over the output of non-metric MDS.

```
## initial  value 15.456070
## iter   5 value 12.100801
## iter  10 value 11.313626
## final  value 11.258762
## converged
```

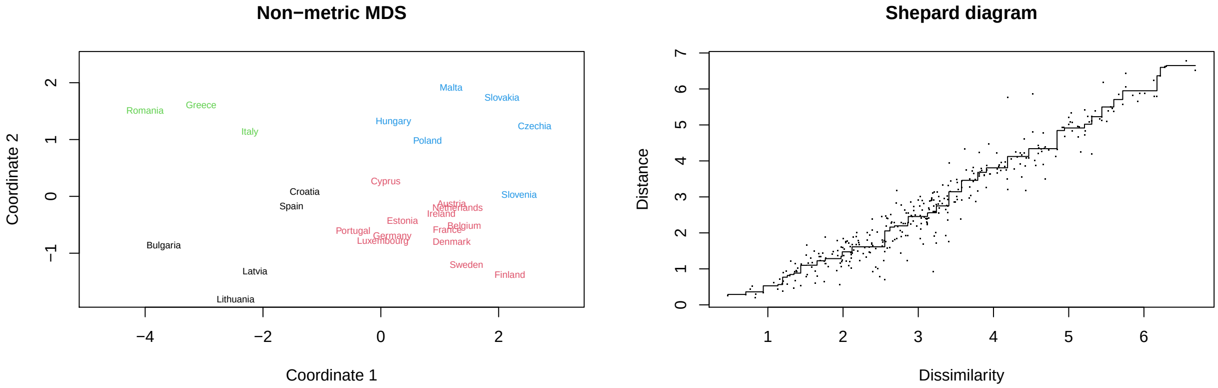

Figure 4: Non-metric MDS

As can be seen from Figure 4 the results are not remarkably different from the previous finding. This time Croatia and Spain happen to be incorporated in the other cluster including Latvia, Bulgaria, and Lithuania. The Shepard diagram helps to assess the goodness of fit. Ideally all points should fall on a monotonic line. The scatter plot shows that EU countries that are close in the input space tend to be close in the output space. Overall, the clusters are fairly robust to the changes in models we have performed (classical and non-metric MDS).

Part b: Canonical Correlation Analysis

(i) Performing canonical correlation analysis among different types of proteint consumption (1st group) and percentage of the workforce by industry (2nd group) in EU countries

A canonical correlation analysis was performed on the following two groups of variables: sectors and dietary. This type of analysis allows to understand the relationships between two sets of variables.

The canonical correlations obtained are:

```
## [1] 0.98997889 0.96451127 0.91495684 0.85120774 0.77008456 0.57073484 0.33326979
## [8] 0.12281991 0.02074961
```

(ii) Commenting on the results, the methodology and package functions used with the help of figures/tables.

The analysis was performed using the `cc()` function from the `CCA` package which performs the analysis over two data matrices. The results are plotted using graphical outputs for canonical correlation analysis available in the `CCA` package. The plot `plt.cc` shows how the variables of the two sets are related in the unit circle.

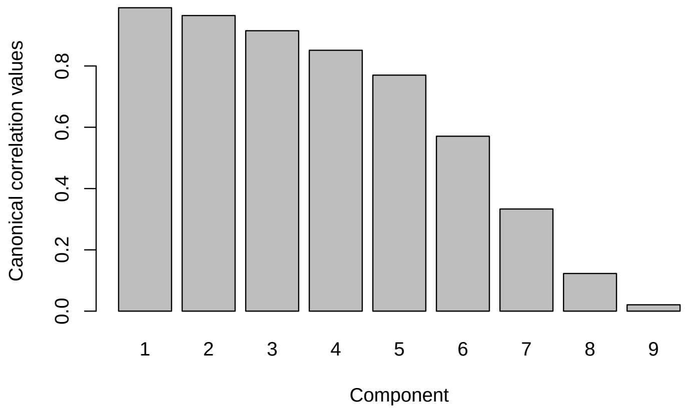

Figure 5: Canonical correlations

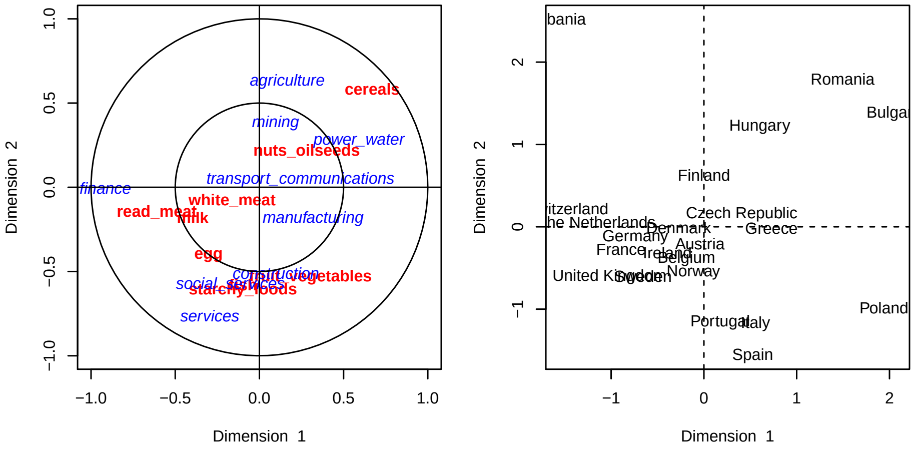

Figure 6: Graphical output of CCA

Interestingly, Figure 6 shows that protein consumption from red meat is closer to the percentage distribution of workforce employed in the finance sector. On the other hand, protein consumption from cereals has stronger associations with agriculture. Finally, protein consumption from fruits and vegetables, fish, and starchy foods have stronger associations with social services and services in general.

We then proceed to test the canonical dimensions by first specifying the number of variables in the second and first set and the number of observations. We then calculate p-values using the F-approximations.

```
## Wilks' Lambda, using F-approximation (Rao's F):
##                  stat      approx df1      df2    p.value
## 1 to 9:  1.497661e-05 1.968305272  81 34.81913 0.01400819
## 2 to 9:  7.510166e-04 1.369510908  64 35.33047 0.15644429
## 3 to 9:  1.077220e-02 1.025923768  49 34.88353 0.47468186
## 4 to 9:  6.614640e-02 0.796633720  36 33.50004 0.74830939
## 5 to 9:  2.401434e-01 0.584642336  25 31.22060 0.91403499
## 6 to 9:  5.900768e-01 0.331387986  16 28.13308 0.98813111
## 7 to 9:  8.751450e-01 0.153262027   9 24.48798 0.99694523
## 8 to 9:  9.844912e-01 0.043151730   4 22.00000 0.99620252
## 9 to 9:  9.995695e-01 0.005168784   1 12.00000 0.94387043
```

The first test of the canonical dimensions shows that all the 9 dimensions are significant. The other tests are not significant and they test the significance of dimensions 2 to 9, 3 to 9 and so on until the 9th dimension.