Project 2

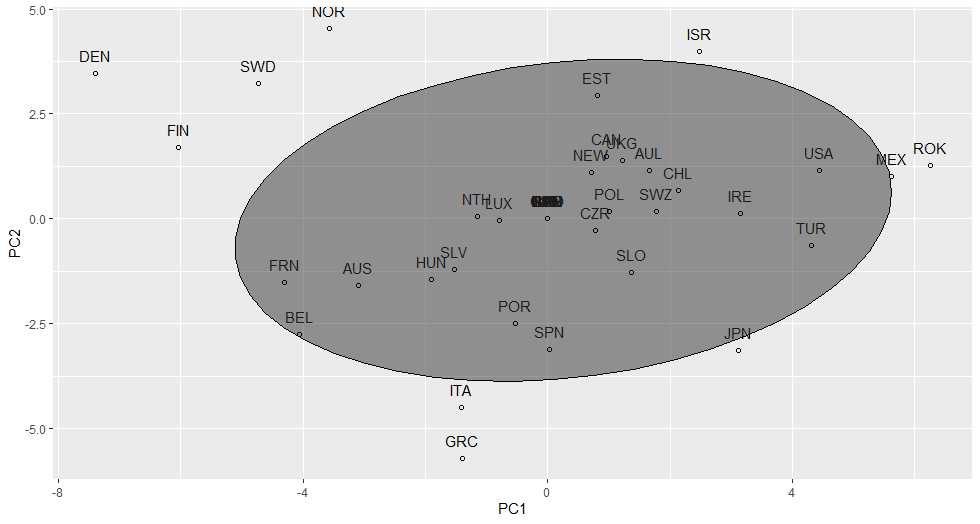
Digital Strategies for Social Science

For this assignment I am working with the Measuring Socialism data set which is a data set compiled from multiple sources. The data is cross-sectional and captures different aspects of governance from 46 countries. More specifically, it includes 243 variables spanning over areas of government taxation, spending, regulation and organizational resources. The set was compiled in order to spur “systematic, better-specified empirical assessment of which countries are more “capitalist”, and how their fates have diverged from more “socialist” countries”. (Cohen and van der Naald, 2019 p.1)

In this assignment I have trimmed down the data set. I am working with 43 countries and ~30 variables. Three countries were dropped because they could not be identified. The data set include all of the Nordic countries except Iceland. The variables I am working with is public expenditure and saving indicators. These variables essentially capture how much money the government spend, towards what purpose the money was spent and which level of government disbursed these expenditures. The variables are treated as numeric variables. In consideration to NAs, I tried out both the missMDA package and column mean imputation. The missMDA package is a more fine-grained way of imputing NAs when doing PCA. However, I decided to use column mean imputation since the data became easier for me to handle and I do not think that the two different methods created significantly different results. Since the data set includes most Nordic countries, my driving research question is whether Sweden will cluster together with their neighbors based on government expenditure and how to best describe Sweden’s dimensional placement. The first part of the question has been studied before. A paper from MIT by Andersen, Holmström et al. finds that “The Nordic countries tend to create a cluster of their own along many dimensions.” (Andersen, Holmström et al. 2007 p. 14). They also state that this cluster tends to be distinctly different compared to other European countries or elsewhere.

I will be utilizing PCA and cluster analysis to answer this question. PCA reduces the dimensionality of the properties in our data based on its variance. The variance retained by each principal component is measured by the eigenvalues. (Abdi and Williams, 2010) We can then group observations together based of this in order to see patterns in our data and assign group memberships to each data point accordingly. PCA is suitable here because we are dealing with numeric expenditure variables and it is generally useful when variables within data sets are highly correlated. We normalize the data when we are doing PCA. This means that we do not supply the original values for the variables. Instead, the PCA is conducted on centered and scaled variable values. This reformats the values to z-scores. There are two benefits to this. The first is that the scale of the variables do not bias the resulting principal components. Secondly, the relationship between our variables stays the same. The inertia explained by my first component is ~.26 and ~.14 for my second component. Therefore, ~.40 of the cumulative variation is explained by the first two eigenvalues. See appendix 1 for scree-plot.

I utilized the ggplot package in R to visualize the observations along the two principal component dimensions. The visualization is overlayed by a 95% confidence ellipse that mark out the majority of observations. Plot 1 indicate that the Nordic countries form a cluster but I need to map this more thoroughly to be sure.

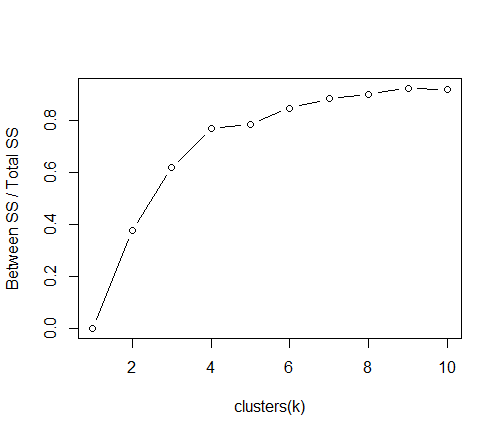


*Plot 1. Observations plotted to the two components*

I look at the correlation between the original variables and the principal components. The output of this correlation is interpreted as Pearson’s R coefficients meaning that values range from 1 to -1, where 0 represent a no linear relationship.

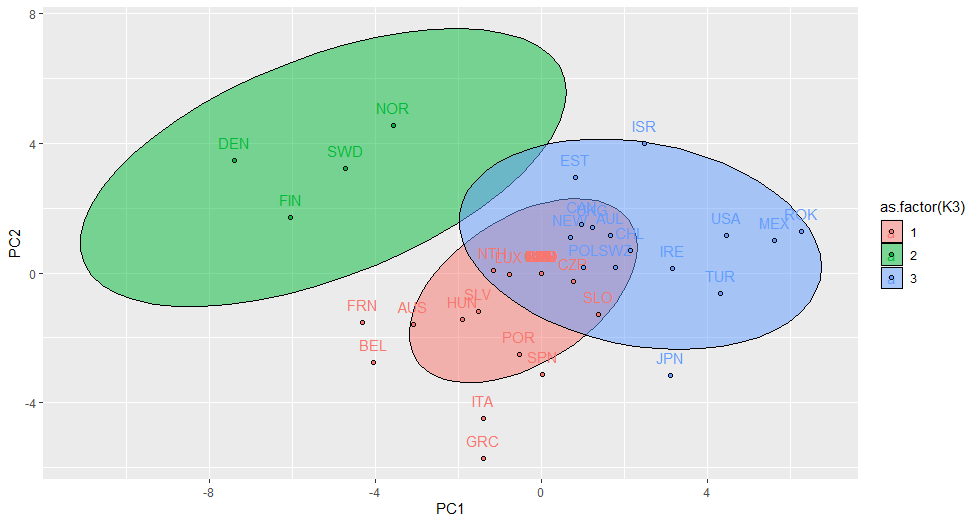
|  |  |  |
| --- | --- | --- |
|  | PRINCIPAL COMPONENT 1 | PRINCIPAL COMPONENT 2 |
| Population | .450 | -.053 |
| Exp. total | -.888 | -.283 |
| Exp. consumption | -.782 | .157 |
| Exp. Military | .292 | .345 |
| Exp. Military. wdi | .242 | .236 |
| Order | .250 | -.439 |
| Exp. econaffairs | -.242 | -.515 |
| Exp. environment | .004 | -.512 |
| Exp. housing | .080 | -.118 |
| Exp. recreation | -.518 | .317 |
| Exp. education | -.325 | .596 |
| Exp. health | -.358 | .054 |
| Exp. socprot | -.810 | -.235 |
| Exp. genservices | -.484 | -.480 |
| Ginv | -.151 | .117 |
| Revenues | -.941 | -.057 |
| Gfinworth | -.201 | .704 |
| Gout | -.118 | .393 |
| Gemp | -.668 | .496 |
| Tariffs | .402 | .176 |
| Soc. oldage | -.649 | -.497 |
| Soc. survivors | -.384 | -.718 |
| Soc. incap | -.736 | .359 |
| Soc. health | -.137 | -.025 |
| Soc. family | -.673 | .391 |
| Soc. labmkt | -.797 | .147 |
| Soc.unempl | -.368 | -.365 |
| Soc. housing | -.309 | .304 |
| Soc.other | -.309 | .433 |
| Soc. total | -.819 | -.218 |

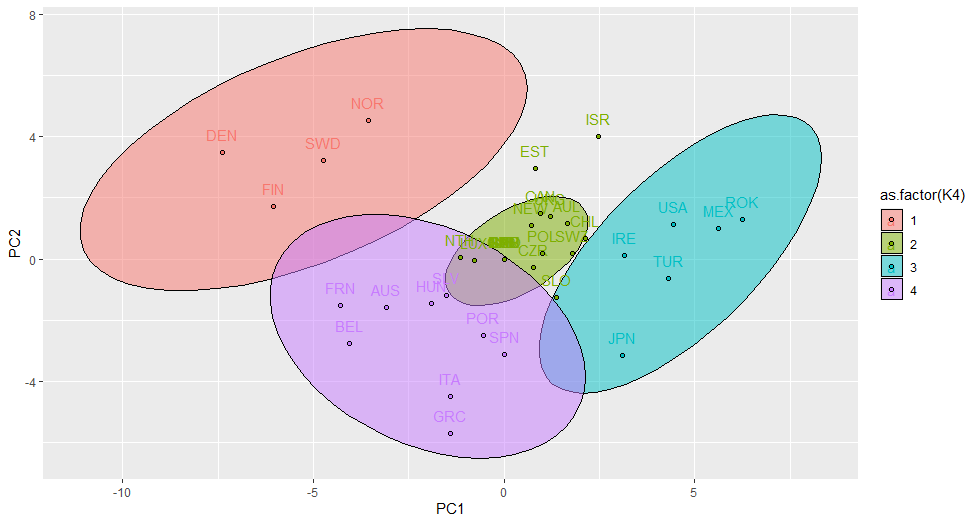
I marked out some coefficients that gives an indication of the difference between PC1 and PC2. Yellow markings indicate coefficients that have large cos2 for PC1. (Abdi and Williams, 2010) See appendix 3 for full visualization of cos2 over the components. For instance, the correlation between Exp. Consumption, which is a measure of the government spending on goods and services used to satisfy the needs of citizens on an individual or collective level, and PC1 is large and negative. A concrete example of such a spending would be the Swedish ROT and RUT deduction (Skatteverket). Generally, as the value for an observation increases along PC1, Exp. Consumption decreases and vice versa for PC2. Coefficients marked in green have a relatively large cos2 for PC2. For instance, as the values for an observation increases for PC2, the total value of the government’s financial assets minus the total value of its outstanding liabilities will go up.

I move on to clustering to investigate underlying group patterns indicated in plot 1. I start with k-Means clustering. k-Means identifies centers in the data based on a specified k parameter. Each observations distance is measured to the k-Means clusters and sorted according to shortest distance. To determine the k parameter, I fitted models with values for k between 1 and 10 and assessed these. The result from this is displayed in plot 2. In plot 2, x-axis show number of clusters (k) and y-axis indicate ratio for between sum of squares and total sum of squares. Reading from the plot I think k should be set to either 3 or 4 so I plotted both for closer visual inspection.

*Plot 2. Optimal k plot*

*Plot 3. k-Means with 3 clusters*

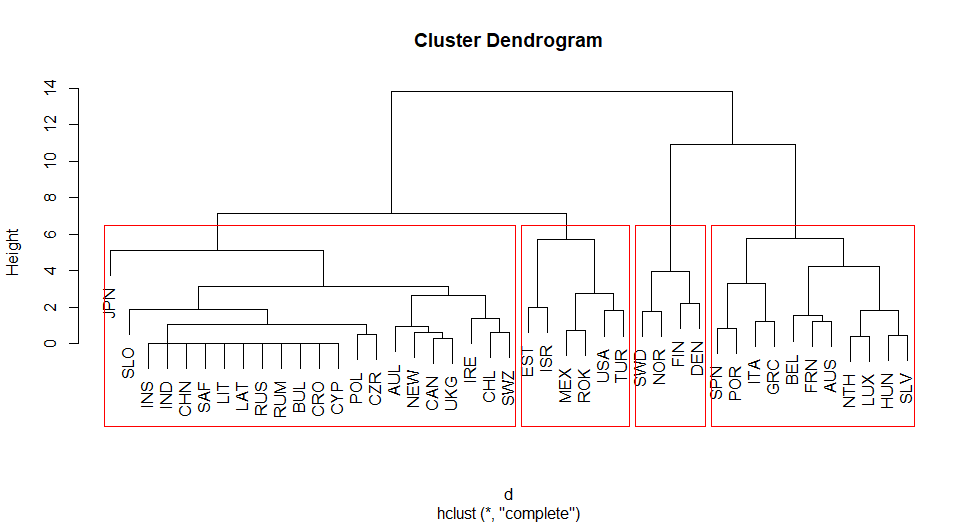




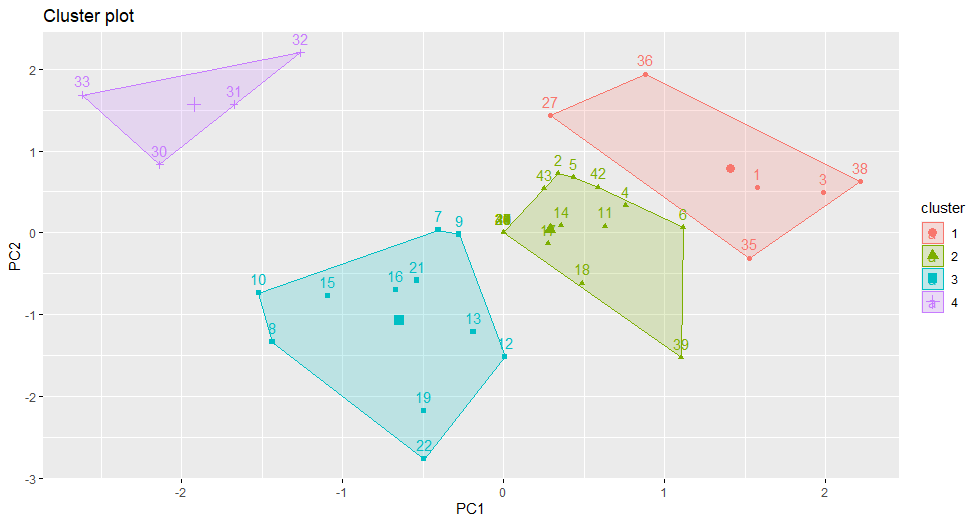
*Plot 4. k-Means with 4 clusters*

Looking at the plots, I believe the most appropriate value for k is 4 which is visualized in plot 4. Most of the observations are captured within the confidence ellipses when k is set to 4. There is some overlap between clusters 4 and 2, yet I think all clusters are reasonably distinguished and separate more nicely compared to when k is set to 3 (plot 3). Clusters 1 and 3 are distinctly different. In cluster 1 we have the Nordic countries and in cluster 2 we have a mix of highly advanced capitalist countries like Japan, USA and Republic of Korea and emerging market economies like Turkey and Mexico. Cluster 4 include southern and central European countries. Cluster 2, although hard to discern by looking, include many eastern European countries as well as Chile, New Zealand and Australia.

I conduct agglomerative hierarchical clustering to see if this will yield a similar cluster output and to better discern the relationships between the clusters.



*Plot 5. Hierarchical clustering output*

k is again set to 4 and the grouping output is quite similar to the k-Means output although a difference can be seen in the way that Japan and Slovenia have been assigned cluster membership. The length of the branches is the Euclidean distance between two points in multidimensional space and reflects the relative difference between the observations. For instance, looking at the large cluster we find a sub-node that contains commonwealth countries like Australia and Canada. The branches in this sub-node are short which informs us that these countries are closely plotted on the PC dimensions and in turn are similar in terms of their governmental expenditure. Comparing plot 4 to 6, we have quite similar clustering using the k-Means and AGNES method.

*Plot 6. Hierarchical cluster output on PC dimensions*

Going back to my initial research question, we can conclude that Sweden is clustered together with other Nordic countries in both k-Means and AGNES method. We can best describe Sweden’s dimensional placement as having a relatively high PC2 score and low PC1 score. Casually speaking, this entails being a rich country that spends a lot of money on education and labor market programs while also stimulating the economy by subsidizing consumption expenditures. There is a cluster forming of Nordic countries in this region of the principal component dimensions. An interpretation of this is that we have identified a region of the PC dimensions that corresponds to the government expenditures of large welfare states. It is an isolate cluster that separates well from any of the other clusters over both methods and an interpretation of this is that the government expenditures associated to the Nordic welfare models are quite unique. The results are in line with the previous literature. Looking at the hierarchical clustering branches in plot 5, we see that Sweden and Norway are more similar to each other than Finland and Denmark. Compared to the southern/central European cluster, advanced capitalist/emerging economies cluster and especially the “mixed-bag” cluster that ranges from commonwealth countries to post-Soviet states, the Nordic cluster is the most geographically homogenous cluster I would say.

This data is cross-sectional and of course government expenditures change over time. It would have been interesting to see if the dimensional placement of Sweden changes over time and by how much, particularly in consideration to which party is in power. This data set was collected between 2015-2017 which means that it was collected during a Social-democratic/left-wing regime in Sweden. How would our results look like over a period of right-wing governance? Also, how would the results change during times of extraneous shocks, like the 2007 financial crisis or COVID-19 pandemic?

**REFERENCE LIST**

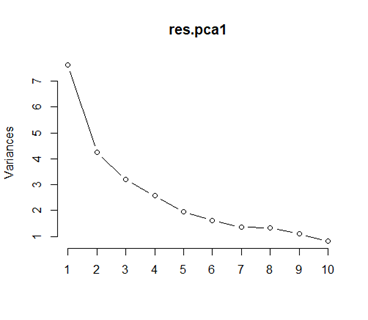
Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews. Computational Statistics, 2(4), 433-459. <https://doi.org/10.1002/wics.101>

Andersen, Torben M. & Holmström, Bengt & Honkapohja, Seppo & Korkman, Sixten & Söderström Hans Tson, & Vartiainen, Juhana,. "[**The Nordic Model. Embracing globalization and sharing risks**](https://ideas.repec.org/b/rif/bbooks/232.html)," [ETLA B](https://ideas.repec.org/s/rif/bbooks.html), The Research Institute of the Finnish Economy, number 232, 01-2015.

Cohen, Joseph Nathan and Joseph van der Naald. 2019. “Measuring Socialism Data Set, version 1.0.” Data Set. Retrieved (osf.io/4g37h).

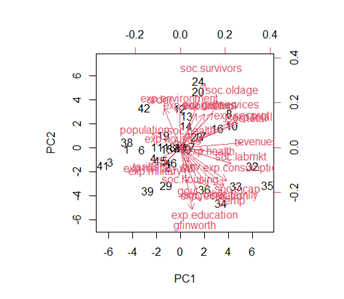
Skatteverket. (9June 2021) *ROT and RUT work*. https://www.skatteverket.se/servicelankar/otherlanguages/inenglish/businessesandemployers/startingandrunningaswedishbusiness/declaringtaxesbusinesses/rotandrutwork.4.8dcbbe4142d38302d793f.html

**APPENDIX 1**

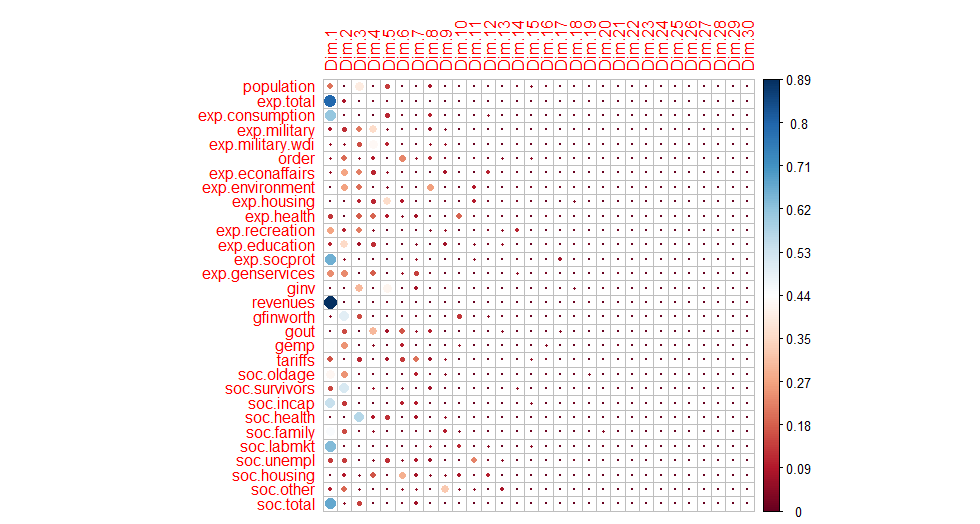
****

*Plot 7. Scree-plot – Elbow at 2 PCs*

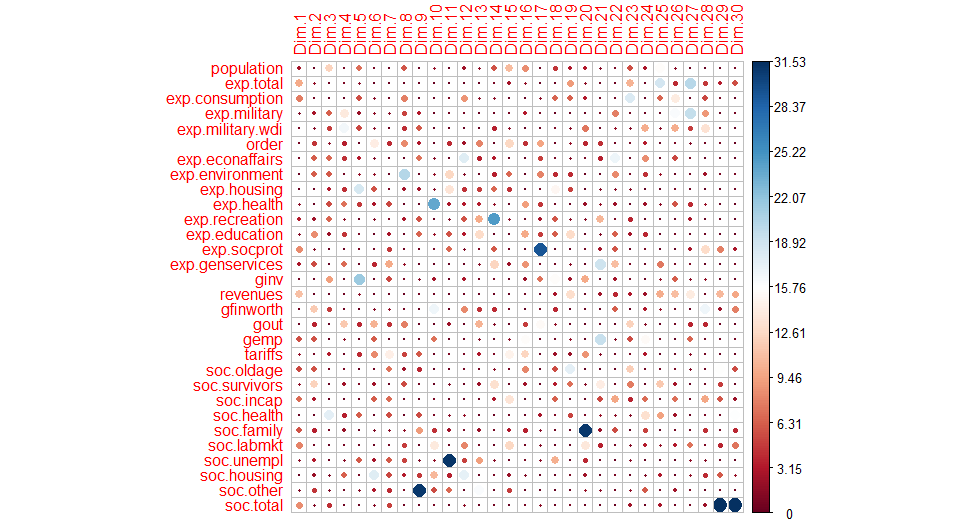
**APPENDIX 2**

****

*Plot 8. Biplot*

**APPENDIX 3**

*Plot 9. Cos2 of variables*

**APPENDIX 4**

*Plot 10. Contribution of variables to PCs*