Bachelor Thesis Business Analytics and Quantitative Marketing

A look at the factorial k-means method

Supervisor: C. Cavicchia
Second assessor: M. van de Velden.
Date definitive version: 4 July 2020
Bachelor Econometrics and Operations Research
Erasmus University Rotterdam

• Joeri Jansen •
Student 426258

Abstract

This thesis uses the methods and data from Vichi and Kiers (2001) to study the factorial k-means method. The factorial k-means methods is applied to a data set, and this data is later used as a baseline in some experiments that test what happens if some values from the baseline case change. In the end, the advantages of factorial k-means will be listed and advise about the choice of data and starting values for the algorithm will be given, based on the experiments.

Contents

1 Introduction

When dealing with data containing an abundance of information, one could get easily overwhelmed by the sheer volume of the data. What possibly could help in the interpretation of the information is the notion of clusters, a group of data points that share a particular property. Clustering helps with the understanding of the data, and is therefore useful when presenting results, as it allows a certain generalization. One of the first people who implemented clustering in the discussion of their results was Driver and Kroeber (1932), and clustering has seen applications in various types of data since. As clustering is used to differentiate between data points, an important question is how different data point should be to no longer belong in the same cluster. One could make a manual distinction between the data points, but this becomes difficult to visualize when the data point contain more characteristics and therefore reach higher dimensions.

A method for data clustering is the so called 'tandem analysis' (named so by Hubert and Arabie (1994)), which first does a PCA (principal component analysis, mentioned in Wold, Esbensen, and Geladi (1987)) on the data set and then uses an algorithm on the scores of the results of the PCA to determine to optimal clustering of the data. However, it could be unwise to use this method as it could lead to a skewed view of the data due to the fact that PCA could identify certain properties of the data set that do not contribute to the clustering structure (De Soete & Carroll, 1994). Vichi and Kiers (2001) propose an other method called 'factorial k-means' which should perform better than the tandem analysis.

This thesis will focus on the factorial k-means method as written in Vichi and Kiers (2001). It will replicate some results from Vichi and Kiers (2001), update the data used and look what the addition of new data points or the changing of starting parameters will do to the results of the factorial k-means method.

2 Literature review

This section will focus more on the work that previously has been done in the field of clustering and factorial k-means.

2.1 Clustering

Clustering uses the principle that data points have certain similarities and differences (or dissimilarities) from each other. Hansen and Jaumard (1997) conclude that most cluster methods have eight steps. First, the sample size is determined. Second, each entity in the sample gets certain characteristics, resulting in a N times p matrix called $\mathbf{X}$, where N is the number of entities in the sample and p is the number of characteristic of the data. Third, the dissimilarities (or similarities) within matrix $\mathbf{X}$ are computed, after which constraints for the clusters are chosen. Next, criteria for the separation of the clusters are implemented. Using the previous two steps, an algorithm is written to calculate the clusters. The next step is using the algorithm on the dissimilarities found in $\mathbf{X}$ to obtain the clusters. In the last step, the results of the algorithm are studied to describe the clusters and find statistic information.

According to Mirkin (1998), the most commonly used clustering methods are k-means (Hartigan & Wong, 1979), hierarchic clustering (Johnson, 1967) and seriation Johnson Jr (1968). K-means begins with picking a set of cluster centroids, and tries to form clusters around these centroids using minimal distance rule, while also updating the centroids regularly. Hierarchic clustering constructs are based on a certain hierarchy. Two schools of thought exist in this subject, as both bottom-to-top methods and top-to-bottom methods are proposes to construct clusters. Seriation uses ordering rules to add items one by one to the clusters, where the clusters get updated each time an item gets added.

2.2 K-means

As discussed in the previous paragraph, one of the methods used for clustering is the so-called k-means method. K-means is based on the study of the so-called clustering error, the sum of all euclidean distances between data points and the chosen cluster centers (or centroids). The cluster centers start at arbitrarily chosen points in the vector space, and are updated and moved with each iteration of the clustering algorithm to minimize the clustering error. Herein lies it biggest weakness however, the original chosen cluster points have a big impact on the final results, and badly chose starting positions will lead to a bad results for the algorithm. Multiple run with different starting points are therefore advised (Likas, Vlassis, & Verbeek, 2003).

The k-means method for clustering is dependent on one piece of information, the number of clusters you want to find. Kodinariya and Makwana (2013) discusses 6 methods one can use to determine a good number of clusters. The simplest method is an arbitrary one, the number of clusters (k) is equal to the square root of $N/2$ ($k{=}\sqrt{N/2}$). A more time consuming method is the elbow method, where different values for k are tested. This method relies on using each value of k for the algorithm, and finding the 'cost' for each run. At a certain k, the cost should drop significantly, which is called the elbow.

2.3 Factorial k-means

Factorial k-means is an extension on the k-means algorithm. Factorial k-means proposes a reduction in the variables, which is based on the principle that certain variables of the data are less important to the clustering structure. Where normal k-means only works with matrix $\mathbf{X}$, factorial k-means also works with loading matrix $\mathbf{A}$, which contain information to which extend the variables p express the clustering structure of clusters k. With this revelation, factorial k-means constructs a loss function (later mention in section 4) and minimizes the sum of the squared distances between centroids in the reduced vector space and the projected data points (Timmerman, Ceulemans, Kiers, & Vichi, 2010).

Factorial k-means is often compared to reduced k-means, primarily because both methods try to form clusters while also reducing the vector space. Factorial k-means was developed more recently than reduced k-means, primarily as a reaction to this method, because reduced k-means does not work in all cases and factorial k-means improves in certain fields where reduced k-means struggles (Terada, 2015). Factorial k-means overall has one major advantage, it uses only one objective function, which is preferred as the sequential use of two functions can interfere with the cluster structure and therefore find sub-optimal results. Factorial k-means and also reduced k-means where created as a reaction to the tandem analysis. Tandem analysis had a major advantage compared to other methods at the time, namely that it produced stable and relatively quick results. Unlike factorial k-means, tandem analysis has two objective functions, factorial k-means is therefore considered to work better, as it has the same advantages as tandem analysis while removing one of its disadvantages (Tortora, Palumbo, & Gettler Summa, 2011).

3 Data

In this section, the data that will be used shall be briefly discussed. Both data from the replication of the work of Vichi and Kiers (2001) will be mentioned, as the updated data.

3.1 The 'old' data

Vichi and Kiers (2001) uses the data of 20 primarily European countries in a demonstration of the factorial k-means analysis. The countries have information about their Gross domestic product (or GDP), their leading indicator, the unemployment rate, the interest rate for 3 months, the trade balance as percentage of GDP and the net national savings as percentage of GDP. All data is in terms of percentage change from the previous year, where the current year is set to 1999. The table with this data set can be found in Table 4 in section A.2.

3.2 The 'new' data

The new data will update the data of Vichi and Kiers (2001) to more current values. This is done by extracting data for 1999 and 2017 (2017 is chosen because it is recent, but not so recent that some data has not yet been collected or is an estimate). All values are retrieved from the OECD database, the following enumeration will provide the appropriate search terms and/or deviations from the norm.

1. **GDP**, found with the search term 'Gross Domestic Product (GDP)', gives the GDP in US Dollars per capita.

2. **Leading indicator**, found with the search term 'Composite leading indicator (CLI)', gives the amplitude adjusted composite leading indicator with a long term average set to 100; because yearly data was not available, a comparison will be made between the data of September 1999 and September 2017.

3. **Unemployment rate**, found with the search term 'Unemployment rate)', gives the unemployment rate in % of the total labor force of a country; data starts at 2003, because earlier years do not include certain countries; the unemployment rate of Switzerland is posted only for years after 2009, so it is set to 4.12% (which is the unemployment rate of Switzerland in 2003 according to Macrotrends LLC).

4. **Interest rate**, found with the search term 'Long-term interest rates', gives the interest rate of government bonds who mature in 10 years; data starts at 2002 because earlier years do not include certain countries.

5. **Trade balance**, found with the search term 'Current Account Balance', gives the current account balance per year in percentage of the national GDP; data starts at 2006, because earlier years do not include certain countries.

6. **Net national savings**, found with the search term 'Saving Rate', gives the saving rate in percentage of GDP.

After the data from both 2017 and a previous year has been collected, the percentile increase (or decrease) for each category is calculated by subtracting the past value from the present value times

100 dived by the absolute value of the past value.
The data consist of 28 countries, the original 20 countries from the 'old' data and the Czech Republic, Hungary, Ireland, Israel, Korea (not mentioned which, but all data found indicates that it is South Korea), Poland, Russia and the Slovak Republic.
In Table 1, the results are displayed for all countries, all countries that were also included in the data of Vichi and Kiers (2001) are highlighted in black. The GDP will be referred to as 'GDP', the leading indicator as 'LI', the unemployment rate as 'Unemp', the interest rate as 'IR', the trade balance as 'AB' and the net national savings as 'SR'. All countries are represented by their 'land code', the full name can be found in Table 3 in section A.1.

Table 1: The percentage change for the economic performance indicators for each of the 28 countries

Country	GDP	LI	Unemp	IR	AB	SR
AUS	89,301	-0,501	-5,645	-54,839	56,430	-10,672
AUT	97,970	0,299	28,323	-88,250	-52,897	15,768
BEL	99,379	-0,115	-13,257	-85,511	-35,841	-40,738
CAN	74,680	0,746	-16,282	-66,288	-294,120	-41,244
CHE	100,383	-0,419	-62,995	-102,253	-55,183	-31,327
CZE	151,260	0,659	-61,123	-79,887	167,801	30,879
DEU	99,967	1,050	7,898	-93,361	35,707	84,797
DNK	106,490	-0,078	49,966	-90,572	133,011	101,104
ESP	98,504	-0,874	-4,242	-68,576	130,252	-32,228
FIN	91,711	0,367	16,135	-89,025	-119,339	-60,130
FRA	83,664	0,166	-12,814	-83,327	-386,637	-47,424
GBR	88,367	-0,001	119,422	-74,750	-25,212	-116,824
GRC	57,538	-0,113	-29,177	16,707	83,546	-19649,882
HUN	171,578	-0,409	41,728	-58,191	132,136	1549,840
IRL	189,720	-0,179	-60,638	-84,110	109,170	-25,443
ISR	72,939	-1,636	29,278	-79,323	-43,194	27,970
ITA	63,206	0,772	-46,593	-58,028	273,842	-52,336
JPN	61,379	-0,105	3,756	-95,910	8,180	-22,819
KOR	143,804	-4,111	1,121	-65,345	2497,372	-3,581
MEX	95,992	-0,666	31,455	-28,412	-384,622	25,641
NLD	89,106	0,492	2,927	-89,324	18,891	13,187
NOR	105,994	0,431	-75,090	-74,351	-71,388	21,635
POL	194,350	0,269	41,588	-53,506	101,873	54,927
PRT	86,735	1,039	-39,566	-39,031	112,662	-82,842
RUS	309,808	-1,330	-53,683	-43,172	-77,163	-9,704
SVK	188,495	3,259	18,340	-86,782	74,940	1431,126
SWE	93,063	0,563	-27,399	-87,649	-62,531	19,313
USA	73,978	-0,950	16,435	-49,467	61,394	-59,245

4 Methodology

This section will explain the factorial k-means algorithm, with a simplified version of the original 1999 data as a demonstration, namely, the non-European countries (being Australia, Canada, Japan, Mexico and the United States).

4.1 Initialization

First, we determine the number of objects (N) and the number of variables per object (p). In our example N=5 and p=6. The data matrix $\mathbf{X}$ is of size (N x p) and consists of all the data we want to include in the clustering algorithm. In our example,

$$\mathbf{X} = \begin{bmatrix} 4.8 & 8.4 & 8.1 & 5.32 & 0.70 & 4.70 \\ 3.2 & 2.5 & 8.4 & 5.02 & 1.60 & 5.20 \\ 0.1 & 5.4 & 4.2 & 0.74 & 1.20 & 15.10 \\ 2.3 & 5.6 & 3.2 & 20.99 & 0.00 & 12.70 \\ 4.1 & 1.4 & 4.5 & 5.59 & -1.40 & 7.00 \end{bmatrix}.$$

Secondly, we want to determine the number of clusters (k), which we set to 2 in this example. With k set, we construct matrix $\mathbf{U}^{*\top}$, which is a matrix of size (N x k) with binary variables that determine to which cluster each object belongs in the initialization faze. Say we initialize the algorithm that the data exists of clusters which contain either English speaking countries or non-English speaking countries,

$$\mathbf{U}^{*\top} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix},$$

because we initialize with Australia, Canada and the United States belonging to cluster 1, and Japan and Mexico belonging to cluster 2. The sum of each row in $\mathbf{U}^*$ is always 1, as objects can only ever belong to one cluster. We also construct matrix $\mathbf{A}$ of size (p x m), where m is the number of components to which the variables are reduced. $\mathbf{A}$ contains information to what degree variables are expressing the clustering structure. $\mathbf{A}$ is restricted by $\mathbf{A}^\top\mathbf{A}= \mathbf{I}_m$. If we use m=2, a possible form of $\mathbf{A}$ is

$$\mathbf{A} = \begin{bmatrix} 0.1 & 0 \\ 0.5 & 0 \\ 0.5 & 0.7 \\ -0.7 & 0.5 \\ 0 & 0.1 \\ 0 & 0.5 \end{bmatrix}.$$

With m=2, we say that the clustering structure exist in a m-dimensional (so 2-dimensional) space. In this example of $\mathbf{A}$, GDP, LI, Unemp and IR are associated with the first dimension (with LI having a bigger effect than GDP and IR having overall the greatest effect on this particular dimension); Unemp, IR, AB and SR are associated with the second dimension. This particular version of $\mathbf{A}$ however is not realistic and probably not even a good starting point, it is meant as a demonstration of the form of $\mathbf{A}$ only, a better form of $\mathbf{A}$ would contain non zero elements in all of its cells. Both $\mathbf{A}$ and $\mathbf{U}^*$ can be chosen freely, as long as the argumentation makes sense (the current $\mathbf{A}$ does not) and constrains for both matrices are not violated (Timmerman et al., 2010; Vichi & Kiers, 2001).

4.2 Loss function

The factorial k-means model has the following mathematical specification:

$$\mathbf{XAA}^\top = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^\top + \mathbf{E}, \tag{1}$$

where $\mathbf{E}$ is a matrix with error terms. $\bar{\mathbf{Y}}$ can be written as $(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{XA}$, which is a useful decomposition, as equation 1 is minimized by the following least square calculation:

$$F(\mathbf{A},\mathbf{U}) = ||\mathbf{XA} - \mathbf{U}\bar{\mathbf{Y}}||^2 = ||\mathbf{XA} - \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{XA}||^2, \tag{2}$$

where $||\mathbf{XA} - \mathbf{U}\bar{\mathbf{Y}}||^2$ is calculated as being the trace of $(\mathbf{XA} - \mathbf{U}\bar{\mathbf{Y}})^\top(\mathbf{XA} - \mathbf{U}\bar{\mathbf{Y}})$.

4.3 Algorithm

The factorial k-means algorithm exists of an initialization step (explained in section 4.1) and 3 repeating steps. It tries to minimize the loss function F ($= F(\mathbf{A},\mathbf{U})$). With $\mathbf{X}$,$\mathbf{U}^*$ and $\mathbf{A}$ determined (and $\bar{\mathbf{Y}}$ calculated as in section 4.2), the algorithm follows these steps:

1. Step 1: Minimize F with respect to $\mathbf{U}$. This is solved per row of $\mathbf{U}$ individually, and it involves looking for which point in the row set to 1, the F function is minimized.

2. Step 2: Update $\mathbf{A}$ with the newfound $\mathbf{U}$, with $\mathbf{A}$ consisting of the first m eigenvectors of $\mathbf{X}^\top(\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top - \mathbf{I})\mathbf{X}$, where $\mathbf{I}$ is the (N x N) identity matrix and $\mathbf{U}$ is the updated $\mathbf{U}$ from step 1.

3. Step 3: With the new $\mathbf{A}$ and $\mathbf{U}$, calculate F again. If step 1 and 2 gave a decrease to the value of F, apply them again. If F has stayed the same, the algorithm has converged and currently rests at a (local) optimum.

5 Results

This section will discuss the results of applying the factorial k-means algorithm on the data sets discussed in section 3, both the old data from Vichi and Kiers (2001) and the new data imported from the OECD. As a rule of thumb, values higher than 0.3 are seen as the most important values in the **A** matrix, and therefore, the corresponding economic variables will be the only ones talked about in the analysis.

5.1 Old data

We differentiate between unstandardized data and standardized data, the standardized data is re-scaled so that it has a mean of 0 and a standard deviation of 1. The results for the unstandardized data will only be briefly discussed, while the standardized data will have a comparison with the results of Vichi and Kiers (2001).

5.1.1 Unstandardized data

When applying the factorial k-means algorithm to the unstandardized old data with k=3 and m=2, we end up with an F-value of 25.547. The three clusters consist of:

1. Australia, Spain and Japan.
2. Austria, Belgium, Canada, Switzerland, Denmark, Germany, France, Great Britain, Greece and Italy
3. Finland, Mexico, the Netherlands, Norway, Portugal, Sweden and the United States.

5.1.2 Standardized data

When applying the factorial k-means algorithm to the standardized old data with k=3 and m=2, we end up with an F-value of 4.096. The three clusters consist of:

1. Australia, Spain, Great Britain, Greece and Japan.
2. Mexico.
3. Austria, Belgium, Canada, Switzerland, Denmark, Germany, Finland, France, Italy, the Netherlands, Norway, Portugal, Sweden and the United States.

If we compare the results for standardized data with the unstandardized data, we learn one thing about the factorial k-means method, the scale of the data will influence the results. The unstandardized data primarily focuses on high GDP and large values for the leading indicator, while GDP is not really important in the standardized data. The unstandardized data had a different range for all characteristics, with different means and standard deviations. This made it that characteristic with many data point far from the mean could dictate a lot of the clustering structure. Putting all characteristics on the same scale makes it so that the algorithm does not look at the value of the data point, but its deviation from the mean, meaning that every characteristic is judged on the same base.

The current results are different from the results of Vichi and Kiers (2001). Cluster 2 loses both Greece and Portugal to cluster 1 and cluster 3 respectively. Cluster 1 gains Great Britain and Japan from cluster 3, but loses also a lot of countries, to the point that only Australia and Spain stay in their original cluster. This could be due to Vichi and Kiers (2001) ending up with a different matrix **A**, the **A** currently used in the optimal solution gives Mexico a very negative value for the second dimension, something that none of the other countries get. Cluster 1 has only countries with positive values for the first dimension, while cluster 3 has the countries with negative values. In extension, cluster 1 has primarily countries with a high leading indicator, while cluster 3 covers the low leading indicator countries. Cluster 2 is primarily know for its high interest rate.

An other potential reason for the difference is the scaling of the data. Vichi and Kiers (2001) scales the data according as a following to the recommendation of Milligan and Cooper (1988). The problem is that they do not point out which rescaling they use, and Milligan and Cooper (1988) discuss 7 methods, including the method used in this paper they call 'Z_1'. It is likely that they used a different rescaling method, one that fits all data points between 0 and 1. The rescaling used in this paper tends to perform less good as opposed to the preferred method of Milligan and Cooper (1988), but if the data has low coverage or random noise, 'Z_1' is better. If we choose Z_4 instead of Z_1, instead of each data point i in characteristic C being redefined as $\frac{i-mean(C)}{standarddeviation(C)}$, i now gets the following form: $i^* = \frac{i}{maximum(C)-minimum(C)}$. If we use the Z_4 method on our data set and run the algorithm, we see however that the F-value decreases, but the clusters stay the same as in the Z_1 situation.

If we use the proposed **U** from Vichi and Kiers (2001), standardize their data using Z_1 and calculate the **A** according to step 2 of the factorial k-means algorithm, the F-value is 35.405. This makes the current clustering structure with an F-value of 4.096 better. It is unclear why Vichi and Kiers (2001) get different results, but it could be attributed to a few factors. First off all, the current algorithm prefers to place countries in the cluster with the lowers number if two cluster would have given the same results in the F-value in step 1 of the algorithm. This could potentially lead to a situation where the algorithm is technically indifferent between placing a point in either cluster 1 or cluster 3, and then placing the country in cluster 1 due to the construction of the algorithm. Secondly, Vichi and Kiers (2001) might have been stuck in a local optimum due to running the algorithm with too few iterations, all results in this paper where gained using 10000 iterations. And while some cases found the optimal solution in less than 1000 iterations, for most calculations 10000 iterations presented a better solution than 1000 iterations.

5.2 New data

The old data uses information attained before the year 2000. It might be interesting to see what happened when we apply the method to more recent data, so we can see how the clusters have changed between the two time periods (which in turn could give insight in how the global economy has changed since the 2000 and how things as the attack on the twin towers, the 2008 financial crisis or the uncertainty of Brexit has impacted the world economy). Such comparisons will be minimal however, as this thesis looks primarily at the factorial k-mean algorithm and not the change of the economy over time. Also, any conclusions based on only the two periods of time will be uninformative, as the data contains relatively little characteristics and countries, and the time frame of the two time

periods are not equal.
The new data will look to the changes to the variables in the last couple of years, and tries to explain what changes in the clustering structure and what causes this. All data used is standardized with the Z_1 method, with k=3 and m=2.

5.2.1 Comparison to 1999

When we use the data from Table 1 and standardize it, we see that some factors (such as GDP or trade balance) have had drastic changes (with increases or decreases of 100% or more). This could very well have a major impact on the clustering structure. When we run the algorithm on all countries featured in the old date, the following cluster appear:

1. Austria, Belgium, Canada, Switzerland, Denmark, Germany, Finland, France, Great Britain, Japan, Mexico, the Netherlands, Norway and Sweden.
2. Australia, Spain, Italy, Mexico, Portugal and the United States.
3. Greece.

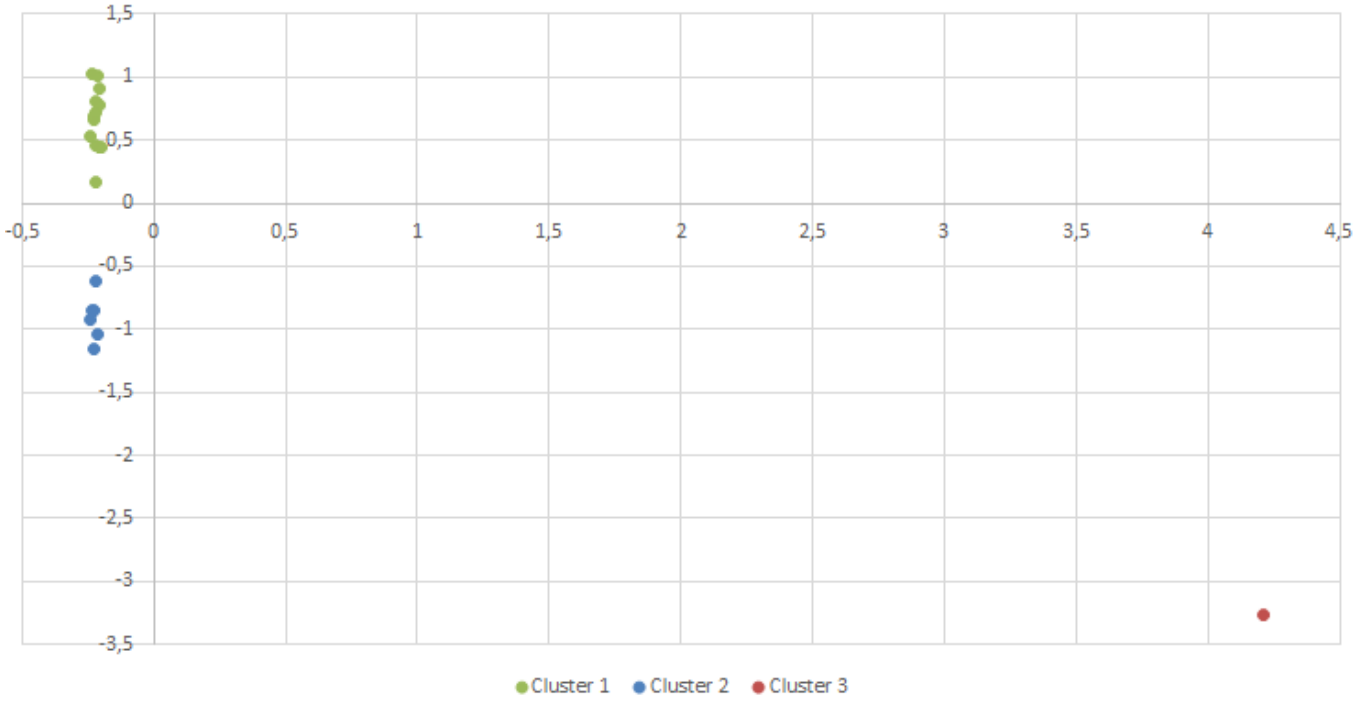

Figure 1: Graphical representation of the current clustering structure

The F-value is 0.929, the 2-dimensional representation of the results is found in Figure 1. Greece is unique, in a sense that is it the only country with a positive value for the first dimension. The second dimension features both positive and negative values, the countries with positive values belong to the first cluster, the countries with a negative value belong to the second cluster. The first dimension is primarily determined with by the value for the net national savings, where a negative net national savings gives the highest score. The second dimension has a same mechanic, but for the interest rate instead of the net national savings.
When we compare this with the results from 1999, we see that GDP is not important anymore for

the clustering structure. Because of this, the clustering structure really changes. Because of the way the data is implemented and the way factorial k-means works, the algorithm primarily tries to find a solution that considers the variables with the biggest variance the most. Because of this, GDP loses its importance in the clustering structure since in the new data, the values for GDP are not that different from each other for the original 20 countries.

5.2.2 Extending the list of countries

But not only do we draw a comparison with 1999, we add a few countries in the list of countries to see what this does to the clustering structure. If the added countries do not alter the averages of the variables of the data, it is expected that the clustering structure stays the same. With this in the back of our head, the clustering structure will most likely change, since Hungary and the Slovak Republic bring relatively high values for net national savings, and the GDP is on average also high of not only these countries but also others. When we run the algorithm on all countries featured in the old date, the following cluster appear:

1. Austria, Belgium, Switzerland, the Czech Republic, Germany, Denmark, Spain, Finland, France, Great Britain, Hungary, Ireland, Israel, Japan, Korea, the Netherlands, Poland, Russia, the Slovak Republic and Sweden.
2. Australia, Canada, Italy, Mexico, Norway, Portugal and the United States
3. Greece.

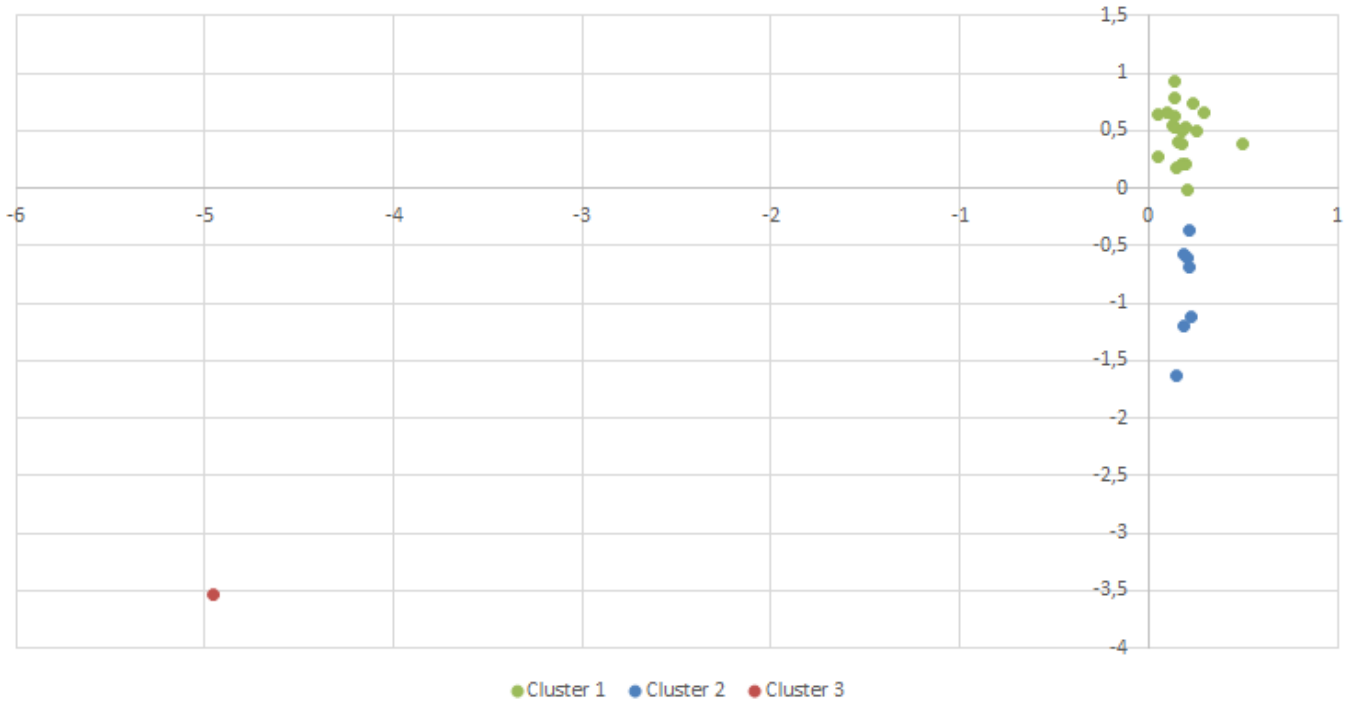

Figure 2: Graphical representation of the current clustering structure

The F-value is 2.398, the 2-dimensional representation of the results is found in Figure 2. The clustering structure is as such because it is once again the only country with a (relatively high) negative value in the first dimension of the reduced vector space. The second cluster is characterized

with a negative value in the second dimension, while the first cluster has only positive values here. The first dimension is primarily associated with the net national savings, making Greece's extreme value for his net national savings the reason it stands alone. The second dimension is primarily positively influenced by the GDP, but negatively influenced by the interest rate. Countries with a (relatively) low interest rate and high GDP are found in the first cluster, while the second cluster hangs more around the average in terms of GDP and interest rate.
In contrast to the last section, GDP once again becomes important in the clustering structure, this is because of the inclusion of countries with relatively high GDP, as hypothesized. This also explains the exchange of countries between the first and second cluster. The second cluster loses Spain when including the 8 other countries, but gains Canada and Norway. Overall, the clustering structure seems to get worse, as the F-value has increased. But since this is explainable by the inclusion of new countries, this is not necessarily worrying. The current clustering structure will be used to compare other structures in the next section when we try to test the limits of the factorial k-means method and how to predict the outcome of the method when applying data.

5.3 Testing the method

In this section, we are going to see how changes impact the results gained in 5.2.2. This in done on 3 different ways: adding or removing clusters, study what qualifies as an outlier in the data and what happens when we add or remove one form the data, and removing characteristics to test the robustness of the clustering structure. Not all **A** matrices from these calculations will not be mentioned in the appendix; when we change the number of clusters, the **A** matrices will be discussed briefly, but not in detail; when we study the outliers, the **A** matrices will not be mentioned as we focus on other results and when we remove characteristics, all **A** matrices will be printed in the appendix as they are the primary form in which we discuss the results.

5.3.1 Adding a cluster

The relatively high F-value could be lowered by introducing an other cluster. If we follow the rule that $k=\sqrt{N/2}$, both $k=3$ and $k=4$ are defensible positions. Since we work in our baseline with $k=3$ and $m=2$, we will draw a comparison with $k=4$ and $m=2$. For these parameters, the clustering structure is as follows:

1. Australia, Austria, Belgium, Switzerland, the Czech Republic, Germany, Denmark, Spain, Finland, Great Britain, Hungary, Ireland, Israel, Italy, Japan, the Netherlands, Poland, Portugal, the Slovak Republic, Sweden and the United States.

2. Canada, France, Mexico, Norway and Russia.

3. Greece.

4. Korea.

The F-value is 0.816. This first cluster is the primarily know to have in both the first and second dimension of the reduced vector space to have positive values. The second cluster has positive values for the the first dimension, but negative values for the second. The third cluster contains only countries who have a negative value in the first dimension of the reduced vector space, while the fourth cluster

covers the extreme case that is Korea, who has a high negative value in the second dimension. Cluster 3 is primarily focused on big net national savings (positive or negative, although in this case negative), cluster 4 has high trade balances (again, positive or negative, but negative in this case), cluster 2 covers countries with relatively high (but not extreme) trade balances and interest rates and cluster 1 contains all other countries.
These results indicate that the inclusion of Korea who has extreme values in one particular feature can bring the structure out of balance. The data in this form contains 2 extreme cases (Korea and Greece), while also containing 26 other countries who do not follow the same pattern. The algorithm prefers to split the 26 countries into 2 groups, and fit one extreme case with it, giving a high F-value in the process because the extreme case who does not get its own cluster will give a large value in the loss function. Given the current results, four clusters instead of three are preferred for this particular data set. However, if either Greece of Korea was deleted from this data set, three clusters would probably give sufficiently good results.

5.3.2 Removing a cluster

Knowing the results of the algorithm for k=3 and k=4, deeper insight might be found if we also test for k=2. With the current data with 2 extreme cases, there are a couple options options for what the clustering structure might look like. The 2 extreme cases might be grouped together, with the other countries in one cluster, or the two countries are split, with the other countries grouped around the extreme case that is most like them. Either way, the results will tell us more about the clustering priorities within this data set. Important to note is that the results will not be very useful, since most methods will discourage k=2 for this particular data set. After we apply the algorithm, the two clusters are:

1. Greece and Korea.

2. Australia, Austria, Belgium, Canada, Switzerland, the Czech Republic, Germany, Denmark, Spain, Finland, France, Great Britain, Hungary, Ireland, Israel, Italy, Japan, Mexico, the Netherlands, Norway, Poland, Portugal, Russia, the Slovak Republic, Sweden and the United States.

The F-value is 8.403. With both Greece and Korea in one cluster, their particular cluster focuses primarily on the net national savings and trade balance. Their cluster is associated with high negative values in the first dimension and positive values in the second dimension, while the second cluster contains all other countries without much further rhyme or reason. Both Denmark and Italy show signs of potentially belonging to the cluster containing Greece and Korea, having both a negative value in the first dimension. However, this is not enough to put them in the first cluster, especially when we consider that their diversion of the average value for trade balance and net national savings is negligible compared to that of Greece and Korea.
When comparing the models with k different from 3, we see that the extreme cases of Greece and Korea determine the majority of the final results. What could be interesting is finding out what the characteristics of countries should be if they want to significantly change the clustering structure. The next section will try to find a metric to determine said characteristic.

5.3.3 Average-median gap

Large chunks of data are usually summarized by either the mean value of the median value. Both metrics are useful in some instances, but generally speaking, the mean is used more frequently. An issue with taking the mean value of a data set is that it gives a distorted result when the range between the minimum and maximum value is large and most data points are centered around the minimum or maximum. Because factorial k-means works with average values, it is susceptible for this issue, the inclusion of Greece and Korea in the data set and its effect on the results should speak for that. So this section will try to find a method that can tell you whether a new value in the data set is extreme or not.

The method to detect such extreme values that I suggest is the absolute difference between the mean and the median value. If this is difference is too big, it tells you that there are data points present in the data set that do not follow the same trend as the other point for a characteristic. And when such a thing gets detected, you have to think about either deleting the data point with the extreme value or introducing a new cluster for this data point, otherwise the result will be distorted. At this point, a concrete decision about what too big means in this context has not been made, primarily because this would depend on the context where this metric would be used. For the current analysis, the variables with the largest gap between mean and median value will be noted as variables of interest and most important in the analysis.

Table 2: Difference between mean and median value

	GDP	LI	Unemp	IR	AB	SR	F-values
With Greece, Korea and the United States	16.57	0.01	2.02	7.30	58.03	593.75	2.400
Without Greece, Korea and the United States	18.19	0.01	0.70	5.71	18.31	121.80	3.344
With the United States	17.56	0.03	2.11	6.38	20.91	115.70	3.444
With Greece and the United States	16.44	0.10	0.49	9.43	22.90	615.50	2.347
With Korea and the United States	17.65	0.05	3.75	6.39	66.50	111.17	3.355

When we look at this mean-median gap for our baseline model in Table 2, we see that both the trade balance and the net national savings have high values. Because the difference for the net national savings is higher, the country that causes this (Greece) has priority in the clustering algorithm. When we remove Greece, Korea and the United States from the data set (the United States gets removed to check what happens when we remove a variable that seems to follow the same trend as the other data point), the differences found at the net national savings and the trade balance decreases considerably, giving reason to believe that the F-value of this data set will be smaller. When we add the United States again, the values for the differences do not change by that much, confirming the suspicion that this particular data point follows the same trend as the other data points and therefore will not interfere with the clustering structure.
Overall, we see that Korea does not follow the same trend as the other countries, therefore, if we include this country, the differences between mean and median increases at the trade balance. This could explain why the F-value is relatively large for the instance with Korea and the United States. However, this does simple way of looking does not explain everything, the instance where neither Korea nor Greece is included has the highest F-value. This can be explained by the fact that standardization

within a sample without outliers leads to relatively high values, but the fact remains that it does not really correspond with expectancy. Further research into outlier detection for multidimensional variables and their effect on the factorial k-means method are therefore advised.

5.3.4 Reducing the vector space

Factorial k-means is based on the principles that all characteristics are equal, but some are more equal than others. For example, if we introduced a characteristic in the data we have been using that has equal values for all countries, the effect on the clustering structure would be 0. Most characteristics do not follow this particular pattern however, but the fact remains that not all characteristics are important in the clustering structure, we have seen this in the data. In the base case with all six characteristics for 28 countries, we defined GDP, interest rate and net national savings as being most important for the clustering structure. With all the other characteristics being less important, it could be interesting to find out what would happened if we delete a characteristic from our data. We will do so with the leading indicator, trade balance and the net national savings.
The clustering structure for data without the leading indicator (F-value is 3.309) is:

1. Australia, Belgium, Canada, Switzerland, the Czech Republic, Germany, Denmark, Finland, France, Great Britain, Ireland, Israel, Japan, the Netherlands, Norway, the Slovak Republic and Sweden.

2. Austria, Spain, Hungary, Italy, Mexico, Poland, Portugal, Russia and the United States.

3. Greece and Korea.

The clustering structure for data without the trade balance (F-value is 5.836) is:

1. Australia, Canada, Mexico, Portugal and the United States.

2. Austria, Belgium, Switzerland, the Czech Republic, Germany, Denmark, Spain, Finland, France, Ireland, Israel, Italy, Japan, Korea, the Netherlands, Norway and Sweden.

3. Great Britain, Greece, Hungary, Poland, Russia and the Slovak Republic.

The clustering structure for data without the net national savings (F-value is 7.682) is:

1. Australia, Austria, Belgium, Canada, Switzerland, the Czech Republic, Germany, Denmark, Spain, Finland, France, Greece, Ireland, Israel, Japan, Mexico, the Netherlands, Norway, Sweden and the United States.

2. Great Britain, Hungary, Poland, Russia and the Slovak Republic.

3. Italy, Korea and Portugal.

When we compare our three new clustering structures to the old one, a few things are worth noting. Greece loses its status as the only country in a cluster, which would be a good sign if it where not for the fact that the F-value increases for all new clustering structures. And even if we remove characteristics with relatively low influence (such as leading indicator or trade balance), the entire $\mathbf{A}$ matrix changes. And these changes are quite unexpected, to the point that the interest rate become important for both dimensions if we remove the leading indicator, while it was only important for the

second dimension in the base case. In this same example, both GDP and net national saving rate lose importance, even though net national saving was the only characteristic important for the first dimension in the base case.

The only interesting thing we can say about the reduction of the vector space is that it does seem that some countries are always bundled together in the cases that we studied. These are: the Czech Republic, Germany, Finland, Ireland, Israel, Japan, the Netherlands and Sweden. Apart from pointing out this trivial fact, reducing the vectors space based on the **A** matrix of our algorithm does not seem like an useful or logical thing to do if it has this impact to the results.

6 Conclusion

Factorial k-means is a clustering method that gives quick and easy to understand results. Comparison between two clustering structure that where obtained using factorial k-means can be easily done based on the F-value of both structures. This F-value is primarily based on the scale used in the data, so caution is advised. The choice for the value of k, or the number of clusters, is important as it can reduce the F-value without impacting the scale of the data. To say that this means that adding more clusters is automatically a good decision is unwise, as to many clusters will give results that are not that useful. Many methods exist to choose the optimal value of k, the easiest to understand being the principle that $k{=}\sqrt{N/2}$, where N is the number of points used to form the entire clustering structure. Outliers are quite an important subject in this clustering analysis, as it impacts both the scale of the data and the clustering structure. Due to the shifts in scale that an outlier can bring, comparing two different clustering structure purely based on F-value gives no determinative information. This thesis can therefore not confidently conclude that adding a data point that looks like an outlier is always a bad idea. More study in the field of outliers is therefore advised. The use of the average-median gap can be useful in this study, but this metric is crude, so giving a first impression is likely all this metric can do.

Reducing the vector space (or number of characteristics of your data) looks interesting, as this allows you to conclude that your data collection process can be simplified. However, this thesis strongly advises against this line of thinking, because removing characteristics from the data does not give predictable results and it could lead to an increase in the F-value (although this can also be caused by the rescale of the data).

A Appendix

A.1 Land codes

Table 3: All land codes used in the paper with their respective country

Code	Land	Code	Land	Code	Land	Code	Land
AUS	Australia	DNK	Denmark	IRL	Ireland	NOR	Norway
AUT	Austria	ESP	Spain	ISR	Israel	POL	Poland
BEL	Belgium	FIN	Finland	ITA	Italy	PRT	Portugal
CAN	Canada	FRA	France	JPN	Japan	RUS	Russia
CHE	Switzerland	GBR	Great Britain	KOR	Korea	SVK	Slovak Republic
CZE	Czech Republic	GRC	Greece	MEX	Mexico	SWE	Sweden
DEU	Germany	HUN	Hungary	NLD	Netherlands	USA	United States

A.2 Tables

Table 4: Table 3 from Vichi and Kiers (2001)

Country	GDP	Leading indicator	Unemployment rate	Interest rate	Trade balance	Net national savings
AUS	4.8	8.4	8.1	5.32	0.70	4.70
AUT	1.1	0.6	4.7	3.84	-0.60	9.40
BEL	1.4	-0.1	9.6	3.64	4.50	12.40
CAN	3.2	2.5	8.4	5.02	1.60	5.20
CHE	1.1	2.1	3.8	1.84	4.40	13.20
DEN	1.0	1.5	5.3	4.08	3.30	5.00
DEU	0.8	-2.0	9.5	3.74	1.50	7.70
ESP	3.6	2.5	19.0	4.83	1.20	9.60
FIN	3.9	-1.0	11.8	3.60	8.80	7.70
FRA	2.3	0.7	11.7	3.69	3.90	7.30
GBR	1.2	4.9	6.4	7.70	-0.50	4.80
GRC	3.2	0.6	10.3	11.70	-8.30	8.00
ITA	0.9	-0.4	12.3	6.08	4.30	8.20
JAP	0.1	5.4	4.2	0.74	1.20	15.10
MEX	2.3	5.6	3.2	20.99	0.00	12.70
NLD	2.9	1.6	4.2	3.69	7.00	15.80
NOR	1.4	0.9	3.3	4.47	7.10	15.10
PRT	2.8	-7.5	4.9	4.84	-8.70	14.00
SWE	4.1	1.1	8.9	4.20	7.00	4.00
USA	4.1	1.4	4.5	5.59	-1.40	7.00

A.3 Starting values and results algorithm

A.3.1 Section 5.1

These matrices are directly pulled from Vichi and Kiers (2001) (Table 3 and 5 respectably). The **A** matrix contains the optimal **A** for the solution of section 5.1.2 .

$$\mathbf{X}=\begin{bmatrix} 4,8 & 8,4 & 8,1 & 5,32 & 0,7 & 4,7 \\ 3,2 & 2,5 & 8,4 & 5,02 & 1,6 & 5,2 \\ 3,9 & -1 & 11,8 & 3,6 & 8,8 & 7,7 \\ 2,3 & 0,7 & 11,7 & 3,69 & 3,9 & 7,3 \\ 3,6 & 2,5 & 19 & 4,83 & 1,2 & 9,6 \\ 4,1 & 1,1 & 8,9 & 4,2 & 7 & 4 \\ 4,1 & 1,4 & 4,5 & 5,59 & -1,4 & 7 \\ 2,9 & 1,6 & 4,2 & 3,69 & 7 & 15,8 \\ 3,2 & 0,6 & 10,3 & 11,7 & -8,3 & 8 \\ 2,3 & 5,6 & 3,2 & 20,99 & 0 & 12,7 \\ 2,8 & -7,5 & 4,9 & 4,84 & -8,7 & 14 \\ 1,1 & 0,6 & 4,7 & 3,84 & -0,6 & 9,4 \\ 1,4 & -0,1 & 9,6 & 3,64 & 4,5 & 12,4 \\ 1 & 1,5 & 5,3 & 4,08 & 3,3 & 5 \\ 0,8 & -2 & 9,5 & 3,74 & 1,5 & 7,7 \\ 0,9 & -0,4 & 12,3 & 6,08 & 4,3 & 8,2 \\ 0,1 & 5,4 & 4,2 & 0,74 & 1,2 & 15,1 \\ 1,4 & 0,9 & 3,3 & 4,47 & 7,1 & 15,1 \\ 1,1 & 2,1 & 3,8 & 1,84 & 4,4 & 13,2 \\ 1,2 & 4,9 & 6,4 & 7,7 & -0,5 & 4,8 \end{bmatrix}, U^* = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{A}=\begin{bmatrix} 0.956 & 0.138 \\ -0.048 & 0.024 \\ -0.008 & 0.087 \\ -0.133 & -0.928 \\ -0.174 & -0.273 \\ 0.192 & -0.194 \end{bmatrix}$$

A.3.2 Section 5.2

These **A** matrices are the results of running the algorithm on the new data for the 20 countries and 28 countries respectably. The first matrices have contain a variable with a -1.000, this is due to the fact that the rounding up was done at 3 decimals. In reality, the value would be $-1+\epsilon$, with ϵ positive and approaching 0. $A^{20} = \begin{bmatrix} 0.007 & 0.033 \\ 0.002 & 0.198 \\ 0.002 & 0.116 \\ -0.007 & -0.932 \\ -0.001 & -0.277 \\ -1.000 & 0.008 \end{bmatrix}$, $A^{28} = \begin{bmatrix} -0.049 & 0.428 \\ -0.056 & -0.229 \\ -0.037 & 0.233 \\ 0.007 & -0.839 \\ -0.030 & -0.079 \\ 0.996 & 0.020 \end{bmatrix}$

A.3.3 Section 5.3.4

These **A** matrices are the results of running the algorithm on the new data without leading indicator, trade balance or net national savings respectably. Dots are used to replace the place of the deleted characteristics.

$$A^{-LI} = \begin{bmatrix} -0.001 & 0.075 \\ ... & ... \\ -0.030 & 0.044 \\ -0.303 & -0.823 \\ -0.767 & -0.086 \\ 0.565 & -0.555 \end{bmatrix}, A^{-AB} = \begin{bmatrix} 0.357 & 0.585 \\ -0.099 & 0.367 \\ 0.140 & 0.494 \\ -0.673 & 0.505 \\ ... & ... \\ -0.624 & -0.157 \end{bmatrix}, A^{-SR} = \begin{bmatrix} 0.574 & -0.521 \\ -0.205 & -0.516 \\ 0.394 & -0.417 \\ -0.331 & -0.368 \\ -0.603 & -0.390 \\ ... & ... \end{bmatrix}$$

A.4 R code

This section contains the code used for the calculations in the paper, the language is R. The code in this form will run for approximately 3 hours.

```
install.packages("matlib")
install.packages("proxy")
install.packages("rstiefel")
# these 3 packages are required to keep the code running

Factorialkmean <- function(a,u,x){
  component1 = x %*% a
  inveteerde = inv(t(u) %*% u)
  sub1 = u %*% inveteerde # calculate U * inverse function
  sub2 = sub1 %*% t(u) # = complete all calculations with U's
  component2 = sub2 %*% component1 # = fully calculate U* Y bar
  whole = (component1-component2)
  F = sum(diag((t(whole) %*% whole))) #calculate the trace
  return(F)
}

mAstart #=the original A matrix
mUstart #=the orignial U matrix
mdata # = the X matrix (does not change in the process)
rescale = scale(mdata) #=standardized version of the X matrix (if needed)

# initialization
X = mdata #choose between either the rescale of the original data
Abestest = mAstart
set.seed(314) #set the random seed, so results stay the same
```

```
for (keer in 1:10000){
  if (keer==1){
    Fstart = Factorialkmean(mAstart,mUstart,X)
    U = as.data.frame(mUstart) #the algorithm continuesly changes U into
        dataframe and back to a matrix
    A = rmf.matrix(mAstart) #make a random orthogonal matrix with the dim
        enions of A
    Fbestest = Fstart
  }
  else{ #randomize the starting value of A and U after 1 iteration of the
      algorithm
    U = U-U
    for (lengte in 1:nrow(U)){
      rando = sample(1:ncol(U),1) #get a random number between 1 and numb
          er of columns
      U[lengte,rando] = 1 # get a random colum in this row to 1, rest is
          still 0
    }
    A = rmf.matrix(mAstart)
  }
  bestU = U
  Fcurrent = Fstart
  Fprev = Fstart+1
  m = ncol(mAstart)
  while (Fprev>Fcurrent){  #run the entire algorithm for this iteration
    for (i in 1:nrow(U)){ # this for-loop will find the best U for this i
        teration
      bestF = Fcurrent
      for (j in 1:ncol(U)){
        U[i,] = list(rep(0, ncol(U))) #reset the column so the next loop
            can work
        U[i,j]=1
        deleter = as.matrix(U)
        check = 1
        while (check<(ncol(deleter)+1)){ #handels the cases where a whole
            row of U contains zeros
          if (sum(deleter[,check])==0){
            gone = TRUE
            deleter = deleter[,-check]
            check = check
          }
          else{
           gone = FALSE
```

```
            check = check+1
          }
        }
        U = as.data.frame(deleter)
        testF = Factorialkmean(A,as.matrix(U),X) #calculate the F-value f
            or the U
        if (ncol(U) < ncol(mUstart)){
          U <- cbind(U,incert)
        }
        if (testF<=bestF){ #check for each column if the current value fo
            r F is better
          bestF=testF
          bestU=U
        }
      }
      U = bestU #update the U for each column to the best u
    }

    #second step of the algorithm, find the A
    Umatrix = as.matrix(U)
    inverse = inv(t(Umatrix) %*% Umatrix)
    front = Umatrix %*% inverse
    back = front %*% t(Umatrix)
    brackets = back-diag(nrow(Umatrix))
    Front = t(X) %*% brackets
    Back = Front %*% X
    eigens = eigen(Back)
    A = eigens$vectors[,1:m] #retrieve the A

    Fprev= Fcurrent
    Fcurrent=Factorialkmean(A,as.matrix(U),X)
  }
  if (Fcurrent<Fbestest){
    Fbestest=Fcurrent
    Ubestest=U
    Abestest=A
  }
  else{
    Fbestest=Fbestest+1-1
  }
}
Fbestest #report the best value for F
Ubestest # report the best value for U
```

A.5 Contribution

The following section contains a paragraph originally in the thesis proposal that explained what the contribution of this paper would be.

This paper will use the updated data to look at both the old data and new data, and compare the results off applying factorial k-means to both data sets, maybe also try to explain shifts between the clusters. After this, it will look what the inclusion of the 8 new countries will do to the data set. I also plan on applying the factorial k-means methods with various values for k (number of clusters/-centroids). I will possibly also take a look what the inclusion of extreme variables will do to the results of applying factorial k-means (extreme variables in this case being variables with characteristics that are unlike the data, so if the data will have the minimum value of a characteristic 5, the average 7 and the maximum 10, the extreme variable will have something like 109). Lastly, if time permits it, I will extract more economic indicators from the OECD database and see what effect that will have on the results.
Basically, my extension will look at the factorial k-means method and see how certain extra variables or choices for initial values change the results of applying the methods compared to a base case (the updated data of the 20 countries originally featured in Vichi and Kiers (2001)). Exact choices for what the differences between the base case and other studies cases will be is not determined at this point in time.

References

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional euclidean space. In *New approaches in classification and data analysis* (pp. 212–219). Springer.

Driver, H. E., & Kroeber, A. L. (1932). *Quantitative expression of cultural relationships* (Vol. 31) (No. 4). University of California Press.

Hansen, P., & Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical programming*, *79*(1-3), 191–215.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Hubert, L., & Arabie, P. (1994). The analysis of proximity matrices through sums of matrices having (anti-) robinson forms. *British journal of mathematical and statistical Psychology*, *47*(1), 1–40.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Johnson Jr, L. (1968). Item seriation as an aid for elementary scale and cluster analysis.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, *1*(6), 90–95.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451–461.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of classification*, *5*(2), 181–204.

Mirkin, B. (1998). Mathematical classification and clustering: From how to what and why. In *Classification, data analysis, and data highways* (pp. 172–181). Springer.

Terada, Y. (2015). Strong consistency of factorial k-means clustering. *Annals of the Institute of Statistical Mathematics*, *67*(2), 335–357.

Timmerman, M. E., Ceulemans, E., Kiers, H. A., & Vichi, M. (2010). Factorial and reduced k-means reconsidered. *Computational Statistics & Data Analysis*, *54*(7), 1858–1871.

Tortora, C., Palumbo, F., & Gettler Summa, M. (2011). *Factorial pd-clustering* (Tech. Rep.). Working paper.

Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*(1), 49–64.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.