Erasmus University

Erasmus School of Economics

Bachelor Thesis Business Analytics and Quantitative Marketing

Parameter Selection for Clustering and Dimension Reduction

Author:
M. Karramass (483819)

Supervisor:
C. Cavicchia
Second Assessor:
M. Velden, van de

July 5, 2020

Abstract

A method for the joint selection of the number of clusters, number of dimensions and α for the clustering and dimension reduction method is proposed here. It can also be used for the factorial k-means, the reduced k-means method and tandem analysis, as these are parts of the general clustering and dimension reduction method. The Average Silhouette Width index and the Calinski-Harabasz pseudo F index are used to assess the model performance for different combinations of parameters. The method is helpful when there is no indication what the parameters should be in advance. This thesis contains a concise explanation of the clustering methods and the two performance measures. Lastly, the performance of the clustering methods are shown on the basis of a simulation study and the parameter selection is illustrated by an economic and a biological example.

Contents

1 Introduction

After the collection of data, researchers and analysts are left with the task to give a valuable interpretation which can be used for different purposes. Large data sets are often incomprehensible, so the focus lies on the reduction of the size of the data. To extract the nontrivial information out of the data set, dimension reduction methods are gaining in popularity (Vichi et al., 2019). Data reduction can be applied on both ends of two-way data. On the one hand, the number of objects can be reduced by various clustering methodologies, e.g. centroid clustering or distribution clustering. On the other hand, dimension decreasing methods like Principal Component Analysis (PCA) (Pearson, 1901) can be used to lower the number of dimensions if it is believed that not all variables have a significant contribution to the interpretation (Vichi & Kiers, 2001).

A frequently chosen method is tandem analysis, described by Arabie & Hubert (1994). Here, the dimensions are first reduced to a pre-selected number of dimensions through PCA. These reduced dimensions are then used to cluster the objects, again in a pre-selected amount of clusters. However, De Soete & Carroll (1994) show that the dimensions with the highest variance chosen in PCA not necessarily contain the information about a possible clustering structure.

To overcome this issue, Vichi et al. (2019) suggest a generalized method of the factorial k-means (FKM) described in Vichi & Kiers (2001), the reduced k-means (RKM) presented in De Soete & Carroll (1994) and the tandem analysis of Arabie & Hubert (1994). This is called the clustering and dimension reduction (CDR) model. An Alternating Least Squares method minimizes its objective function, which depends on the weight of each of the methods in the generalized objective function. Certain parameters have to be chosen in advance. These parameters are the desired number of clusters, number of dimensions and the ratio of the tandem analysis and the FKM method. The RKM method is a combination of these two, right in the middle. The Vichi et al. (2019) propose a method to select these parameters through the Calinski-Harabasz pseudo F index. This method is then only executed to find the optimal ratio. The choice of the number of clusters and dimensions are not further investigated in that paper.

In this thesis, I focus on the question: *"How to jointly select all the parameters of the clustering and dimension reduction method together?"*. I first illustrate the different methodologies on the basis of a simulated example with well-separated clusters. Afterwards, I apply the methods to real data, given the parameter choices of Vichi & Kiers (2001). These results are compared to the results when using a joint selection of the parameters through overall Average Silhouette Width (ASW) index (Rousseeuw, 1987) and the Calinski-Harabasz pseudo F (pF) index (Calinski & Harabasz, 1974).

The motivation is to further develop the usage of FKM and RKM. Big data analysts can use this research for the choice of the parameters when (a mix of) these methods are used. I use different data sets to illustrate the performance of the methods. First, I simulate a data set with a clear clustering structure and randomly generated masking variables, which is similar to what is done in Gordon (1999). Next, I show an economic application with the same data set as in Vichi &

Kiers (2001), which contains macroeconomic data of OECD countries. For a varying number of dimensions, the optimal number of clusters is chosen. Lastly, I use a biological data set concerning the milk composition of different mammals as presented in Spector (1956).

The remainder of this thesis is structured as follows. In Section 2, I discuss relevant literature describing the clustering methods and the performance measures of the clustering structures. Section 3 contains more details about the data that I use in this thesis. In Section 4, I then discuss the methods used for the clustering and the performance measure of the resulting clustering structures. Afterwards, I state and examine my main findings in Section 5. I conclude this thesis with remarks, limitations and possibilities regarding improvement of the research in Section 6.

2 Literature review

The main idea of this research field is to reduce the data with a minimal loss of nontrivial information. An intuitive and popular method is the tandem analysis. It is called tandem analysis because it is a 'tandem' of two different methods, PCA (Pearson, 1901) and k-means (MacQueen, 1967), proposed by Arabie & Hubert (1994). The idea behind this analysis is to use the projection of the objects on the components found in PCA. These projections are then used to cluster the data with the k-means algorithm.

The tandem analysis can have poor performance, as PCA does not necessarily select the dimensions that contribute to the clustering structure most (Vichi & Kiers, 2001). Nevertheless, the idea of reducing the dimensions and the clustering of the objects can still be successful. De Soete & Carroll (1994) present a new clustering method, called the RKM method. Here the dimensions and clustering are found simultaneously, in contrast to the tandem analysis. In this method, the centroids of the clusters are considered and chosen in such a way that the distance between the cluster points and the cluster centroids is minimized.

Later, Vichi & Kiers (2001) present a new method, FKM. This method is similar to RKM, with a difference in the objective function. If there is a lot of variability in the directions orthogonal on the information about the clustering, RKM fails to recover this important information (Vichi & Kiers, 2001). To overcome this, the FKM method considers the projection of the objects on the subspace. Now, the orthogonal distance is taken away when minimizing the distance between the centroids and the objects. On the other hand, if there is a lot of subspace variance, RKM has a better performance (Timmerman et al., 2010).

Vichi et al. (2019) present a mix of FKM, RKM and tandem analysis, a general method named CDR. If there is reason to believe that a mix of the methods has a better performance, these can be mixed through addition of the objective functions of tandem analysis and FKM. They both have a weight between 0 and 1, summing up to 1. If the weight of both the objective functions is equal to 0.5, the CDR model is equal to the RKM model.

Extensive research has been conducted on parameter selection. Markos et al. (2018) show that

both the ASW index as well as the pF index give valuable information on the selection of the number of clusters and number of dimensions. The ASW index measures the similarity of an object to its own cluster in comparison to the other clusters (Rousseeuw, 1987). The pF index is a measure of the ratio of the between-cluster and the within-cluster variance (Calinski & Harabasz, 1974). Both are used as a performance measure of cluster classifications.

3 Data

I use multiple data sets, each for a different purpose. First, the differences between the RKM and FKM are shown through a simulated data set with three well-separated clusters (Gordon, 1999). Secondly, the macroeconomic data about the OECD countries used in Vichi & Kiers (2001) are used to first further describe the performance of FKM, and is afterwards used to illustrate the method of parameter selection. Finally, biological data about the milk composition of 25 mammals illustrate a more advanced application of the joint selection of the parameters for the CDR method. The data about the countries and their macroeconomic variables and mammals and their milk composition are included in Appendix A.

3.1 Simulated data

To illustrate the performance of the RKM and FKM methods, I construct a division of 42 objects in three well-separated clusters (Gordon, 1999). This can be seen in Figure 1. I mask this structure by adding four extra variables. The values of these variables are randomly generated by a normal distribution with mean 0 and standard deviation 6. The centroids of the clusters are on the corners of a equilateral triangle with side length 6. The idea behind this rather trivial example, is to test the ability to recover the masked structure of the clustering methods.

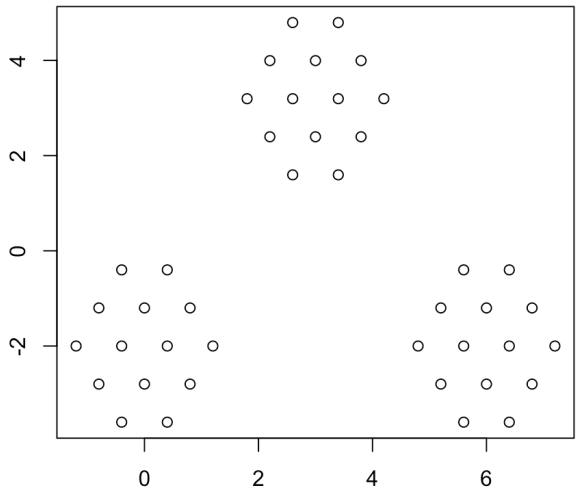

Figure 1: 42 objects divided in three well-separated clusters.

3.2 Macroeconomic data OECD countries

The same macroeconomic data as in Vichi & Kiers (2001) are used for the application of the FKM method in comparison to the outdated tandem analysis. There are 20 countries included in the data set, together with six macroeconomic indicators (from the scenario in 1999). These six indicators are: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB) and Net National Savings (NNS). This data set is downloaded from the *clustrd* package in R, named *macro*.

3.3 Milk composition of mammals

Animals can be subdivided in different classes with similar traits. One of those classes are mammals *(Mammalia)*. These mammals have several things in common, for example that they feed their young with milk. However, the composition of milk differs for the different species. I investigate the constituents of the milk of 25 animals. There are 5 percentages of interest: water, protein, fat, lactose and ash. This data set is downloaded from the *cluster.datasets* package in R, named *all.mammals.milk.1956*.

4 Methodology

In this section I start with introducing the relevant terminology used in the rest of this thesis. Afterwards, I present the technical aspects of the CDR method, which includes tandem analysis, FKM and RKM. I conclude this section with the presentation of the ASW and pF index.

4.1 Terminology

For the remainder of this thesis, $\mathbf{X}$ denotes the $(n \times k)$ data set with n objects and k variables, $\mathbf{U}$ the $(n \times c)$ matrix which indicates the allocation of the object to one of the c clusters, $\mathbf{Y}$ the $(c \times m)$ matrix containing the m dimension coordinates for the cluster centroids, $\mathbf{A}$ the $(k \times m)$ loading matrix with $\mathbf{A}'\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is equal to the identity matrix of order m and $\mathbf{E}$ the $(n \times k)$ residual matrix.

4.2 Reduced *k*-means

The RKM method creates centroids for the clusters in a low-dimensional subspace. The method then minimizes the distance between the objects from the full space and the 'quasi' centroids in the subspace, as stated in De Soete & Carroll (1994). They present the model given in Equation 1 as the model fitted by the RKM model.

$$\mathbf{X} = \mathbf{UYA}' + \mathbf{E}. \tag{1}$$

The loss function that is minimized to obtain the desired result is stated in Equation 2.

$$f_{RKM}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{UYA}'||^2. \tag{2}$$

Here, $||\cdot||$ denotes the Frobenius norm. Through minimizing Equation 2, I obtain values for the loading matrix and the allocation of the objects to the different clusters. In practice, this method is equivalent to maximizing the between variance of the clusters in the reduced space (Vichi et al., 2019).

4.3 Factorial *k*-means

The FKM method projects all the objects on a low-dimensional subspace. From here, it minimizes the distance from the centroids in this subspace to the projections (Vichi & Kiers, 2001). This is unlike the RKM method, where only the centroids are in the subspace. An in-depth comparison between the theoretical performance is out of the scope of this thesis, Timmerman et al. (2010) discuss the similarities and the differences of RKM and FKM in more detail. FKM in essence fits the model in Equation 3 (Vichi & Kiers, 2001).

$$\mathbf{XAA}' = \mathbf{UYA}' + \mathbf{E}. \tag{3}$$

The loss function, which again has to be minimized for the best fit in Equation 3, is given in Equation 4 (Vichi & Kiers, 2001).

$$f_{FKM}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = ||\mathbf{XAA}' - \mathbf{UYA}'||^2 = ||\mathbf{XA} - \mathbf{UY}||^2. \tag{4}$$

I derive the optimal fit of the loading matrix and the allocation of the objects to the clusters through minimization of Equation 4. Effectively, the FKM method minimizes the sum of squared distances of the projected data points to their respective centroids (Vichi et al., 2019).

4.4 Clustering and dimension reduction

The CDR method is a method which generalizes tandem analysis, FKM and RKM. The loss function is given in Equation 5, as stated in Vichi et al. (2019).

$$f_{CDR}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \alpha||\mathbf{X} - \mathbf{XAA}'||^2 + (1-\alpha)||\mathbf{XA} - \mathbf{UY}||^2. \tag{5}$$

In essence, the CDR loss function is a combination of the loss function in tandem analysis and the loss function of FKM. The loss function of FKM is stated in Equation 4. The loss function in tandem analysis is equal to

$$f_{Tandem}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{FA}'||^2, \tag{6}$$

where the optimal F is equal to XA (Vichi et al., 2019).

For $\alpha = 0$, the loss function of the tandem analysis drops out and the loss function is equal to

the FKM loss function. For $\alpha = 1$ the loss function of FKM drops out and that leaves us with the tandem analysis loss function. A remarkable property of the CDR model is that if $\alpha = 0.5$, the loss function coincides with the RKM loss function (Equation 2). This is shown in Equation 7.

$$\begin{aligned} f_{CDR,\alpha=0.5} &= 0.5(||\mathbf{X} - \mathbf{XAA}'||^2 + ||\mathbf{XA} - \mathbf{UY}||^2) \\ &= 0.5(||\mathbf{X}||^2 + 2\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{XA}) + \text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{XA}) - 2\text{tr}(\mathbf{X}'\mathbf{UYA}) + ||\mathbf{UYA}||^2) \\ &= 0.5(||\mathbf{X} - \mathbf{UYA}'||^2) = 0.5 f_{RKM}. \end{aligned} \tag{7}$$

The scalar multiplication of the RKM loss function has no influence on the optimal solution. The model allows for variables to be nominal and/or ordinal, but as it is not used in this thesis, this is omitted in the explanation. For more extensions to the CDR model, see Vichi et al. (2019).

4.5 Alternating Least Squares

Minimizing the loss function is done through the Alternating Least Squares (ALS) algorithm described in Vichi & Kiers (2001) and in Vichi et al. (2019). Here, I present a compact overview of the execution of the ALS algorithm for our purpose.

Initialization For the algorithm to commence, I choose initial values for $\mathbf{A}$, $\mathbf{U}$ and $\mathbf{Y}$. The values for $\mathbf{A}$ and $\mathbf{U}$ can be chosen in a sensible way, or randomly if there is no indication of a good starting point. They do have to satisfy the constraints. Every object has to be allocated to exactly one cluster and $\mathbf{A}'\mathbf{A} = \mathbf{I}$, as stated earlier. The value of $\mathbf{Y}$ follows from the choice of $\mathbf{A}$ and $\mathbf{U}$, as $\mathbf{Y} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}$.

Step 1: Updating U The $\mathbf{U}$ is only present in the FKM part of the CDR loss function. This means that optimizing $\mathbf{U}$ is equal in both the CDR method and the FKM method. The optimal $\mathbf{U}$ minimizes the loss function of the FKM method (Equation 4) with respect to $\mathbf{U}$. For now, we leave the other variables constant. The optimal $\mathbf{U}$ is found through independently considering all elements per row u_{ij} with $i \in \{1, .., n\}, j \in \{1, .., c\}$. We choose $u_{ij} = 1$ if $\min\{f(A, [u_{ij}])\} = \min\{f(A, [u_{ik}]) : k = 1, ..., c\}$ and $u_{ij} = 0$ otherwise.

Step 2: Updating Y and A Now $\mathbf{U}$ is optimized, we are left with the task to optimize $\mathbf{Y}$ and $\mathbf{A}$. As $\mathbf{Y}$ can be expressed in $\mathbf{U}$ and $\mathbf{A}$, we fill in $\mathbf{Y} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}$ and optimize $\mathbf{A}$. In order to do that, we minimize Equation 8.

$$
\begin{aligned}
f_{CDR}(A,U,Y) &= \alpha||\mathbf{X}-\mathbf{XAA}'||^2 + (1-\alpha)||\mathbf{XA}-\mathbf{UY}||^2 \\
&= \alpha||\mathbf{X}||^2 - \alpha\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{XA}) + (1-\alpha)||\mathbf{XA}-\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}||^2 \\
&= \alpha||\mathbf{X}||^2 - \alpha\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{XA}) + (1-\alpha)\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{XA}) - (1-\alpha)\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}) \\
&= \alpha||\mathbf{X}||^2 + (1-2\alpha)\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{XA}) - (1-\alpha)\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{U}'(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}) \\
&= \alpha||\mathbf{X}||^2 + \text{tr}(\mathbf{A}'[(1-2\alpha)\mathbf{X}'\mathbf{X} - (1-\alpha)\mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}]\mathbf{A}) \\
&= \alpha||\mathbf{X}||^2 + \text{tr}(\mathbf{A}'\mathbf{X}'[(1-2\alpha)\mathbf{I}_n - (1-\alpha)\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}']\mathbf{XA})
\end{aligned}
\tag{8}
$$

Note that $\mathbf{I}_n$ is the identity matrix of order n. Effectively, it boils down to setting $\mathbf{A}$ equal to the first m eigenvectors (ordered from high to low) of the matrix $\mathbf{X}'((1-\alpha)\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' - (1-2\alpha)\mathbf{I}_n)\mathbf{X}$ (Vichi et al., 2019).

Repeating these steps result in a monotonically decreasing loss function. There is no insurance that the algorithm ends up in a global optimum. To overcome this, I use 100 different starting values. Vichi & Kiers (2001) state that in practice this is sufficient.

4.6 Parameter selection

Before executing the aforementioned methods, I have to choose the value of certain parameters. These parameters are the number of clusters c, number of dimensions m and value of α. In De Soete & Carroll (1994) and Vichi & Kiers (2001), the values for c and m are fixed without further substantiation. To select these parameters, I propose the ASW index and the pF index. When choosing the parameters which are getting tested, Vichi & Kiers (2001) show that m should not be larger than $c-1$. This because the addition of extra dimensions has no contribution, where it does increase the data size. This can be explained by the fact that one only needs $n-1$ dimensions to display n centroids.

4.6.1 Average Silhouette Width (ASW) index

The ASW index validates the consistency within clusters. It measures the similarity of every object to its own cluster compared to the other clusters. The *Silhouette Coefficient* for object i is defined in Equation 9 (Rousseeuw, 1987).

$$
S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{9}
$$

Here, $a(i)$ is expressed as the average dissimilarity of the i^{th} object to the other objects in the same cluster. Specifically, this is equal to the mean intra-cluster distance. The parameter $b(i)$ shows the mean nearest-cluster distance. A higher ASW index corresponds with a better fit. The values of $S(i)$ lie in the interval $[-1,1]$. The ASW index is then equal to the average value.

4.6.2 Calinski-Harabasz pseudo F (pF) index

The proposed method to choose these parameters according Vichi et al. (2019) is through the pF index (Calinski & Harabasz, 1974). This index provides a measurement of the quality of the subdivision of the objects to the clusters. The statistic is given in Equation 10 (Vichi et al., 2019).

$$pF = \frac{\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A})/d_b}{\text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}))/d_w} \tag{10}$$

The degrees of freedom in Equation 10 are given by $d_b = (c-1)m + (k-m)m$ and $d_w = nk - cm - (k-m)m$. The pF index is in essence equal to the quotient of the between cluster variance and the within cluster variance in XA, respectively weighted by their degrees of freedom. This implies that the maximal index value is optimal. The index value can only be calculated after the partitioning is found, which means that CDR has to be executed multiple times with different parameters in order to compare the performance. The CDR method is computationally efficient, which makes this a minor issue.

5 Results

This section shows the application of the discussed method on three data sets. The results are split into three sections, one for each data set. With the data, I illustrate the strength of the clustering methods and how to select the parameters. Both the macroeconomic and the biological variables are standardized as proposed in Milligan & Cooper (1988). For the execution of the FKM, RKM and CDR methods, I used the R package **clustrd** (Markos et al., 2018) in combination with my own programming. This can be found in Appendix B.

5.1 Clustering structure recovery after masking

To test the performance of the tandem analysis, RKM and FKM, I use a simulated data set. Here, two variables together construct three clear clusters and the remaining four variables are randomly generated from the normal distribution to mask the structure. For all the methods, I fix $c = 3$, and for RKM and FKM. I also fix $m = 2$, as in Vichi & Kiers (2001). As a baseline, I attempt to recover the three well-separated clusters that are masked by four randomly generated variables through k-means. The results are displayed in Figure 2. Here, the original clustering structure is visualized and the displayed objects are labeled with the clustering results of the k-means clustering algorithm.

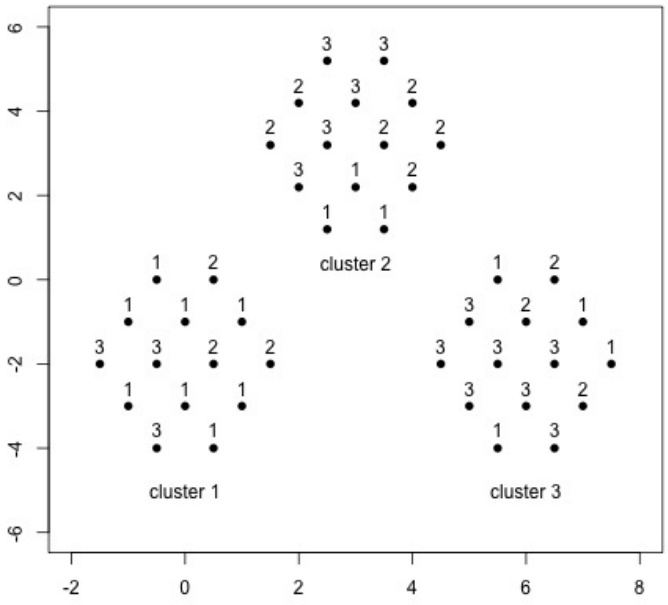

Figure 2: K-means partition of three well-separated clusters, masked by four randomly generated variables, plotted on the underlying clustering structure.

Figure 2 shows that k-means partition is not powerful enough to separate the clustering structure from the masking variables. It is not capable to recover more than half of the original classification. This proves that the k-means algorithm is not fit to analyze data where the underlying structure is masked by randomly generated variables.

5.1.1 Tandem analysis

As an attempt to improve the performance, I can use PCA first. This is equal to the tandem analysis. In Figure 3, I used PCA with a different number of components, varying between 2 and 5. In Figure 3a, the objects are projected on the subspace of the first two principal components, labeled with their original cluster allocation. In the remainder of Figure 3, the underlying clustering structure is plotted and the labels correspond to the classification according to the tandem analysis.

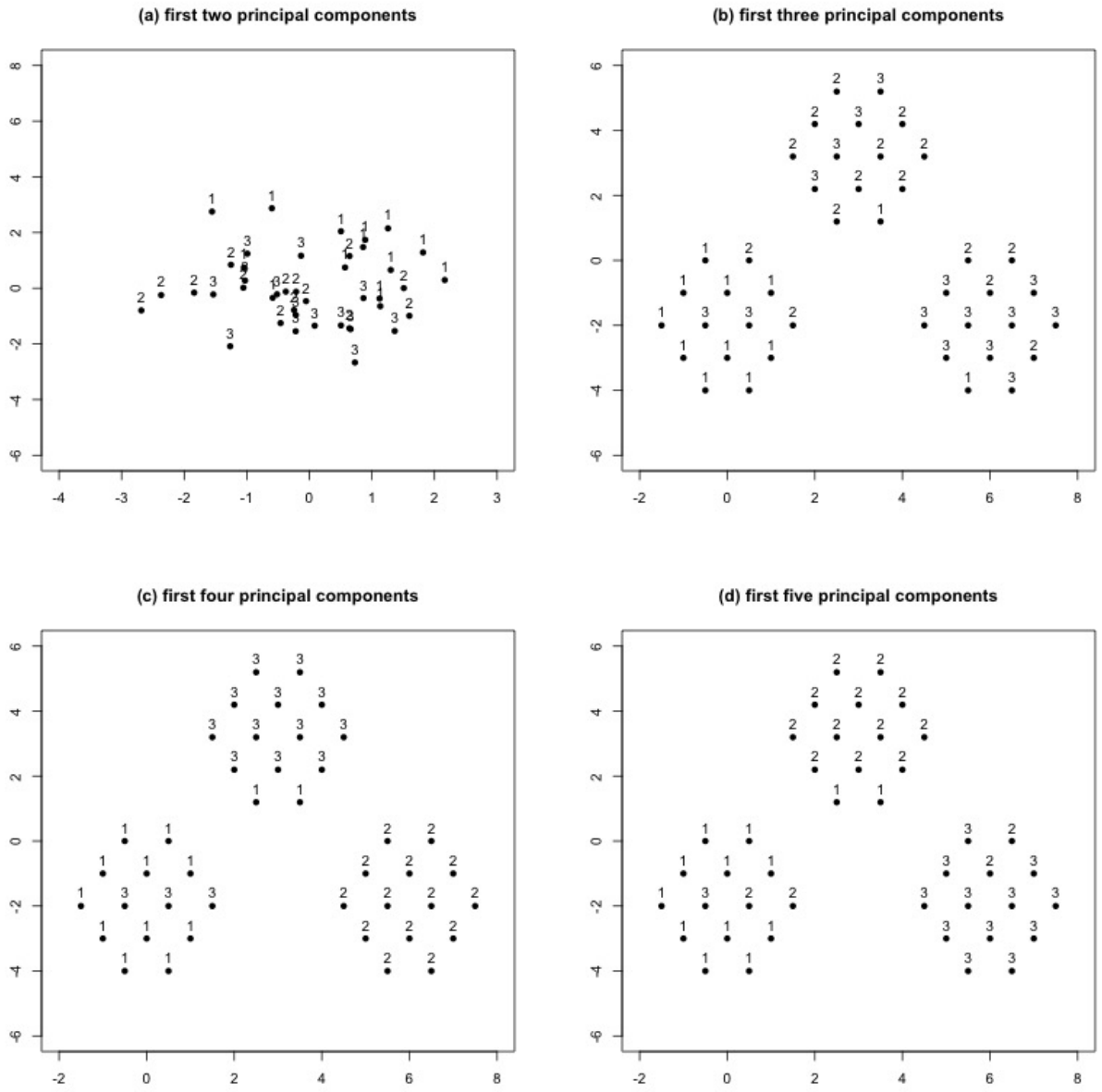

Figure 3: Tandem analysis applied on a varying number of principal components.

It is clear that the first two principal components are unable to find the clustering structure, as there is no clear division of objects visible. This is explained using Table 1. Table 1 also confirms the rest of the information Figure 3 presents. Namely, that the majority of the variance can not be captured by a minimal number of principal components. The third principal component still covers over 10% of the total variance. This can also be seen by the decreasing number of wrongly classified objects when increasing the number of used principal components.

Table 1: Captured variance by each of the principal components.

Component	1	2	3	4	5	6
Eigenvalue	1.2887	1.1619	1.1292	0.8447	0.7684	0.6403
Proportion of Variance	0.2768	0.2250	0.2125	0.1189	0.0984	0.0683
Cumulative Proportion	0.2768	0.5018	0.7143	0.8333	0.9317	1.0000

5.1.2 Reduced k-means

The tandem analysis appears to perform poorly in this application. I turn to the RKM method. This allows the dimensions to adapt to the maximization of the between variance of the clusters. The results of this method are visualized in Figure 4. The labels correspond to the original clustering structure with the first two constructed dimensions on the axes.

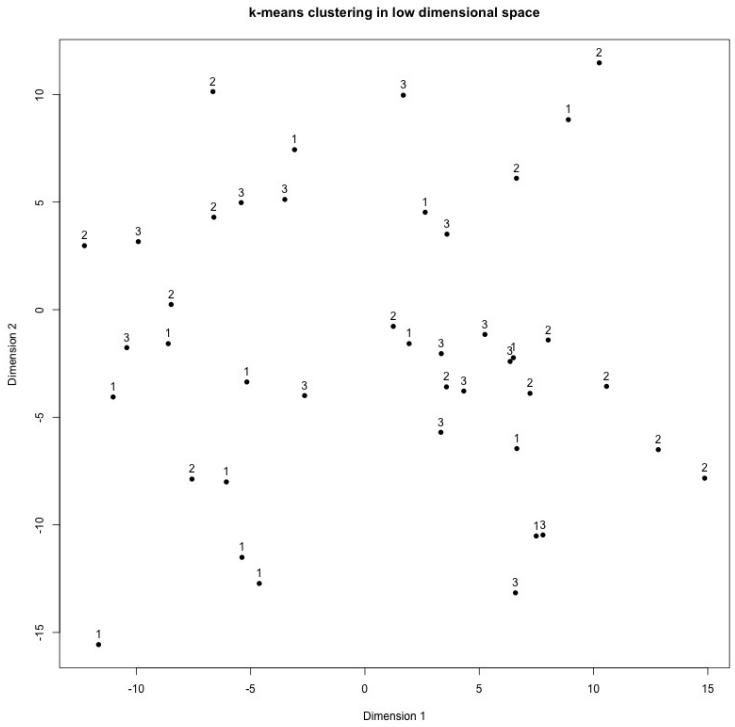

Figure 4: Visualization of objects after reduced k-means. The objects are labeled with their correct cluster classification.

The RKM model also fails to distinctly divide the objects to the desired clusters. There are no clear groups visible and objects with different labels are intertwined. This corresponds with the earlier mentioned technical background of the method. Specifically, the performance of the RKM method deteriorates when the data contains much variability in directions orthogonal to the subspace of the data containing information about the clustering structure. This is indeed the case with the randomly generated variables.

5.1.3 Factorial k-means

Lastly, we look at the FKM model. This model, theoretically, should have a better performance in the circumstance of a clustering structure masked with variables orthogonal on the relevant subspace. The performance can be seen in Figure 5, again with the constructed dimensions on the axes.

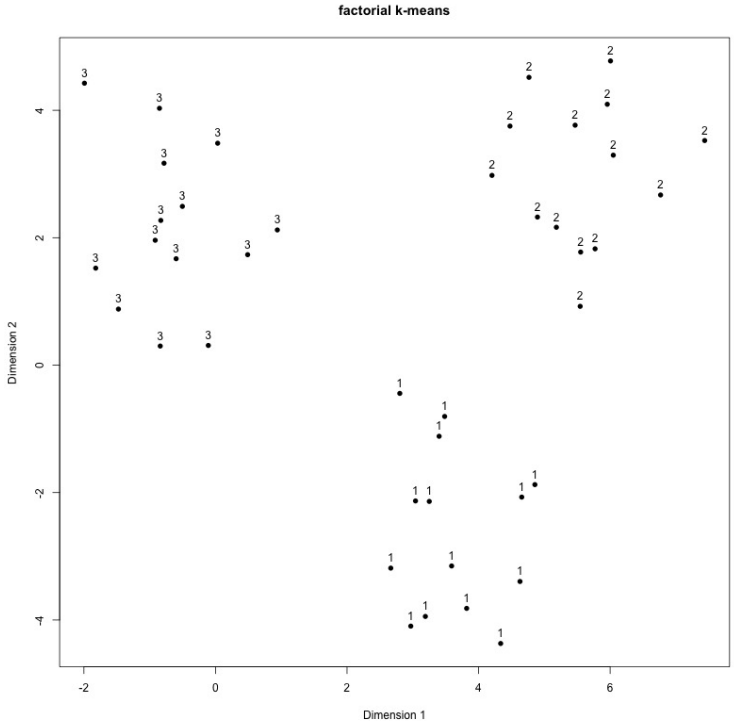

Figure 5: Visualization of objects after factorial k-means. The objects are labeled with their correct cluster classification.

Now we see that FKM is powerful enough to detect the underlying clustering structure. There are three well-divided clumps of objects visible, where every object is classified in the correct cluster. To further investigate the performance of FKM, I look at the correlation between the constructed dimensions and the variables. These can be seen in Table 2.

The first two dimensions correlate highly with the variables that contain the information about the clustering structure. We conclude that FKM indeed is the best performing method when han-

Table 2: Correlation between the constructed dimensions of factorial k-means method and the six variables.

Dimension	1	2
X-coordinate	0.94	-0.35
Y-coordinate	-0.35	-0.93
Masking variable 1	0.03	0.01
Masking variable 2	-0.02	0.05
Masking variable 3	-0.00	0.03
Masking variable 4	0.00	-0.06

dling much variability orthogonal on the subspace of the original variables that define the clustering structure. The excellent performance is conditional on the knowledge about the correct number of clusters. To illustrate, I add an extra cluster and I apply FKM with $c = 3$ and $c = 4$. This is shown in Figure 6.

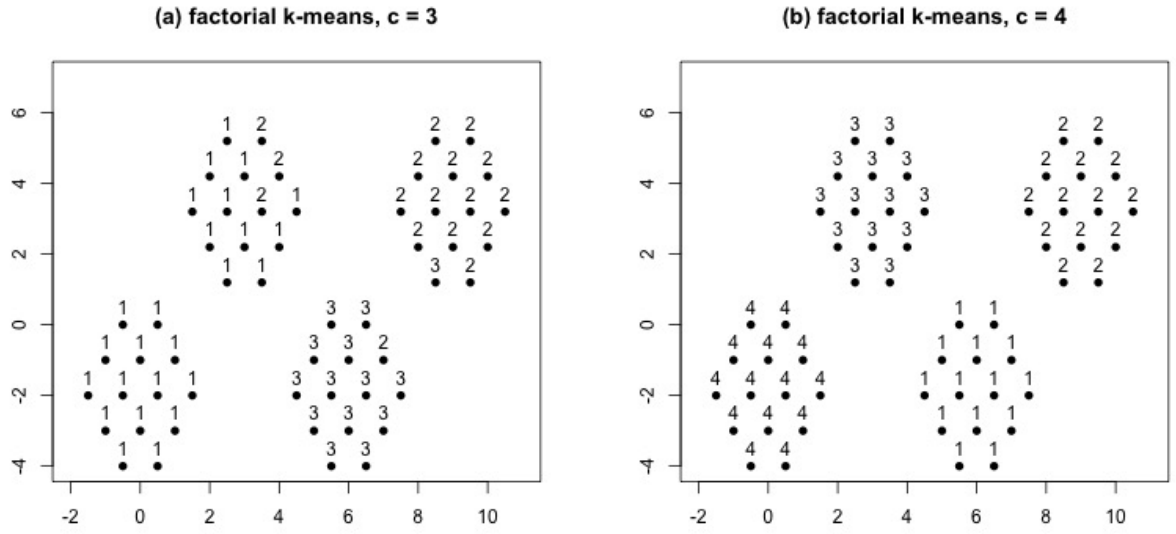

Figure 6: Visualization of objects after factorial k-means. In (a), the method used three clusters and in (b), the method used four clusters.

Not only does FKM miss an important facet of the underlying clustering structure if I use $c = 3$ instead of $c = 4$, namely an incorrect number of clusters, it also classifies multiple objects incorrectly in the clusters it does detect. This issue can be remedied by detecting the correct number of clusters. This can be done with the use of the ASW and/or the pF index. For the structure shown in Figure 6, the index values for both indices are graphically visualized in Figure 7.

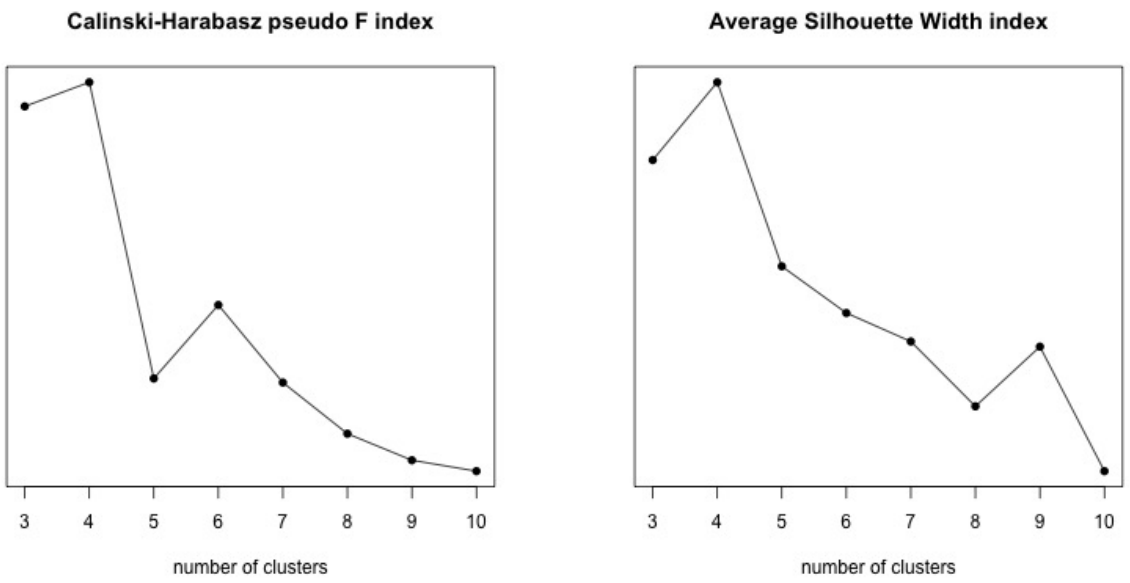

Figure 7: pF index and ASW index for a varying number of clusters with simulated data.

For both indices, the highest index value corresponds with the model with four clusters. This shows that the usage of these indices contribute to an optimal clustering strategy.

5.2 Economic application: macroeconomic scenario 1999

To show an application of the performance of FKM analysis, I exert this method on the macroeconomic scenario of 20 OECD countries in 1999. I first calculate the ASW and the pF indices to determine the number of clusters and dimensions, where Vichi & Kiers (2001) assumes that $c = 3$ and $m = 2$. When the parameters are selected, I cluster the countries and try to retrieve the ground of clustering structure that is detected. Vichi & Kiers (2001) also give the relative performance of FKM in comparison to tandem analysis, where it is evident that the FKM method outperforms the tandem analysis. From a theoretical perspective, the FKM method does not necessarily perform better than the RKM, because it is unknown how the variables relate to each other. For further analysis I use the CDR method, as it is a generalization of the methods. After the selection of the parameters for FKM, I investigate whether the CDR method with a variable α obtains higher index scores for $\alpha \neq 0$, as $\alpha = 0$ results in the FKM method.

5.2.1 Parameter estimation

To determine the values of the c and m, I calculate the ASW and pF indices. I test c for values from 3 up to 10. I do not check higher values, as a higher number of clusters results in insufficient reduction of data. For m, I check 2 and 3 dimensions. This again for sufficient reduction and also to maintain a low number of dimensions for an interpretable visual display. The index scores are shown in Figure 8.

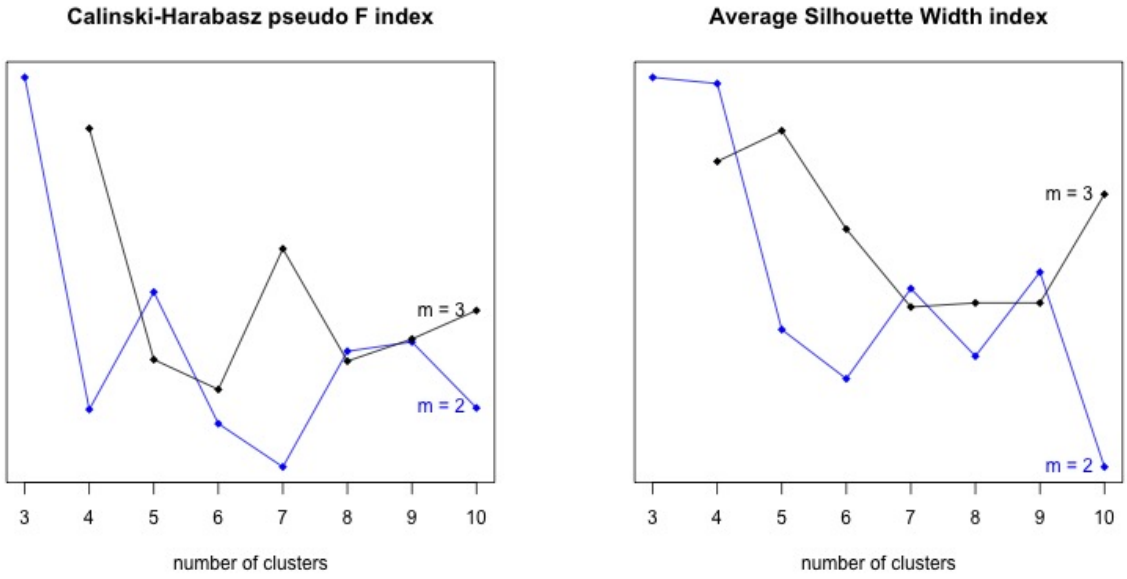

Figure 8: pF index and ASW index for a varying c from 3 to 10 and for $m = 2$ (blue line) to $m = 3$ (black line) dimensions with macroeconomic data of 20 OECD countries from 1999.

Both indices point towards the use of $m = 2$ and $c = 3$. The results of the FKM method with the optimal parameters are shown in Figure 9a. To investigate the underlying clustering structure that the FKM method has detected, I also look at the correlation between the two dimensions and the macroeconomic variables. This is shown in Figure 9b.

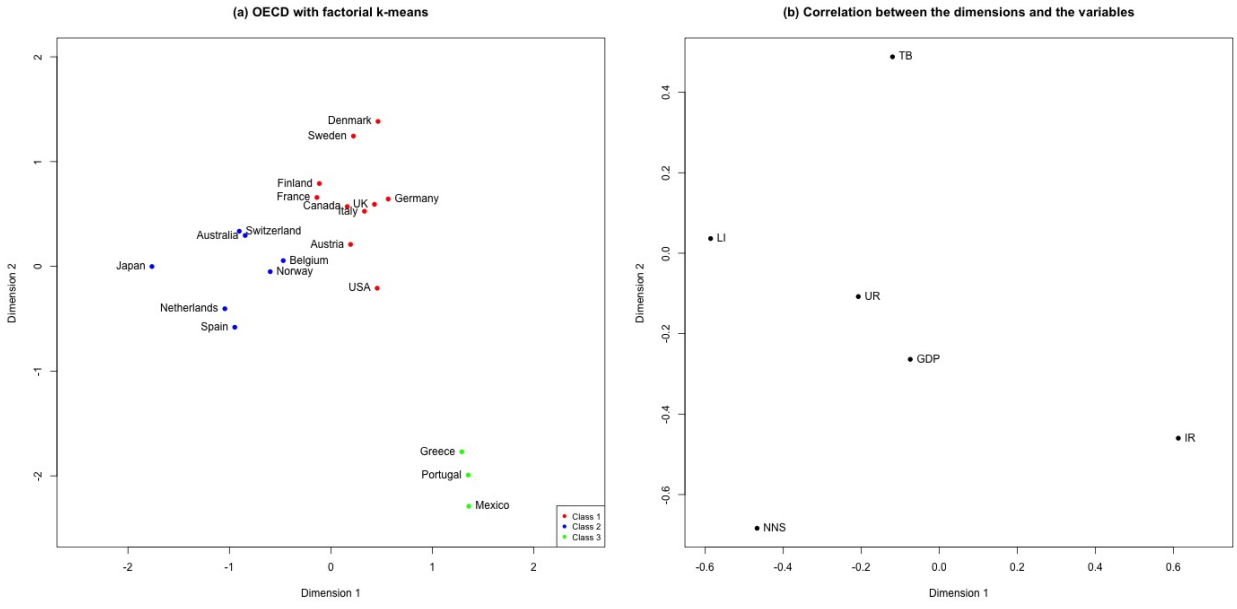

Figure 9: 20 OECD countries divided in three clusters is shown in (a). In (b) the macroeconomic variables are projected onto the plane spanned by the two dimensions.

When we look at Figure 9b, we see that GDP is relatively underrepresented in the correlation with the dimensions, together with the unemployment rate. The leading indicator is heavily correlated with the first dimension. The trade balance is strongly correlated with the second dimension. The net national savings and interest rate are correlated with both. This gives information of which variables are strongly involved in the construction of the dimensions and thus in the detected clustering structure.

5.2.2 Application of CDR

The application of the FKM method on the OECD data as seen earlier is not automatically better than the RKM method. It is proven that the FKM method performs well when the for the clustering structure irrelevant data and the relevant data are orthogonal. In the simulated example, this is obvious. For the macroeconomic data, this is unknown. To assess whether a mix of different methods increases the index values, I evaluate the ASW and pF index for the CDR method for a variable α. The parameters $c = 3$ and $m = 2$ remain unchanged. The results are visualized in Figure 10.

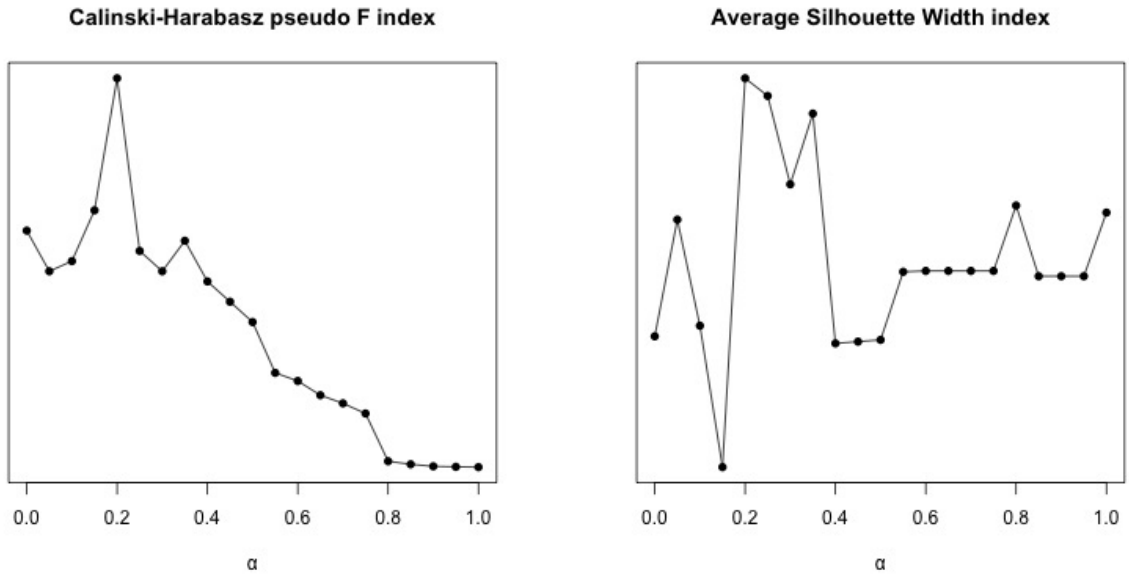

Figure 10: pF index and ASW index for a varying α from 0 to 1 with steps of 0.05 with macroeconomic data of 20 OECD countries from 1999.

As seen in Figure 10, both indices point towards an α of 0.2. As $0.2 \in (0, 0.5)$, this means that the index suggests a mix of the RKM and FKM method.

It now remains interesting whether the importance of the variables drastically changes when I use CDR with $\alpha = 0.2$. Figure 11a shows the resulting structure of the CDR method. In Figure 11b I visualize the correlation between the two dimensions and the macroeconomic variables.

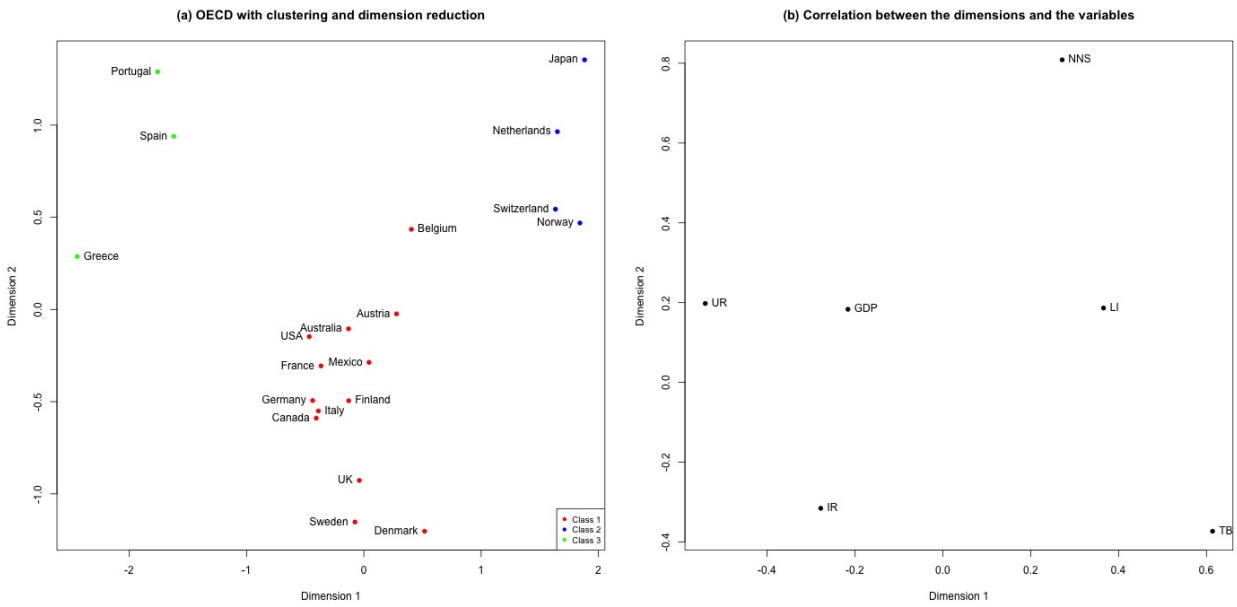

Figure 11: 20 OECD countries divided in three clusters through CDR is shown in (a). In (b) the macroeconomic variables are projected onto the plane spanned by the two dimensions.

If we compare the clustering of FKM in Figure 9a with CDR in Figure 11a, we see some similarities as well as some differences. Portugal and Greece are still close to each other and distant from the majority of the countries. A difference is that Spain is closer to the majority of the countries, where it lies closer to Greece and Portugal when I used FKM. This can be explained by the slight differences in Figure 11b in comparison to Figure 9b. For example, the correlation between the dimensions and the trade balance decreased, where the importance of the unemployment rate increased. Overall, there are no major changes in the importance of the macroeconomic variables. Another remarkable result is that the three clusters are better divided when I use the CDR method. In the FKM method, classes 1 and 2 are close to each other. With the CDR method, we see a clearer division between the countries in different clusters. This was also expected, as higher index values correspond to a better division of the clusters.

5.3 Biological application: milk composition of mammals

It can be interesting to investigate whether a clustering method can detect a sensible clustering structure for animals, based on their milk composition. I expect the animals with similar traits close to each other. For this application, I only use the generalized CDR method, as it includes all the other methods. To investigate whether the CDR model can detect a sensible clustering structure for these animals, I first assess the parameters. I only consider the ASW index for a consistent selection of the parameters. The reason I choose the ASW index and not the pF index, is because the former has a better overall performance (Arbelaitz et al., 2013).

We calculate the ASW index values for α varying between 0.05 and 1 with steps of 0.05, c from 3 to 10 and m for 2 and 3. I restrict the interval for the latter two parameters for the same reasons as stated in Section 5.2.1. The results are shown in Figure 12.

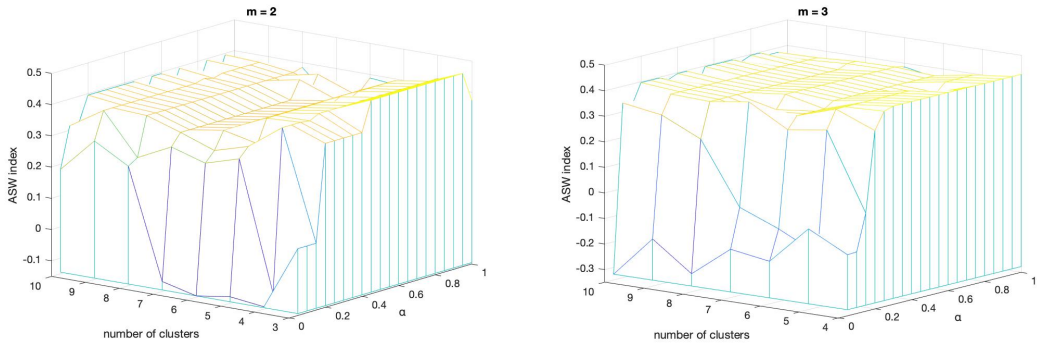

Figure 12: ASW index for the CDR method with a varying α, number of clusters and number of dimensions.

We observe that the maximum index value is equal to 0.496. This value is reached for $m = 2$, $c = 3$ and $\alpha \in [0.45, 0.95]$. This implies that for different values of α the same clustering structure is detected. To find this clustering structure, we choose $\alpha = 0.45$. The detected clustering structure together with the correlation between the dimensions and the variables can be seen in Figure 13.

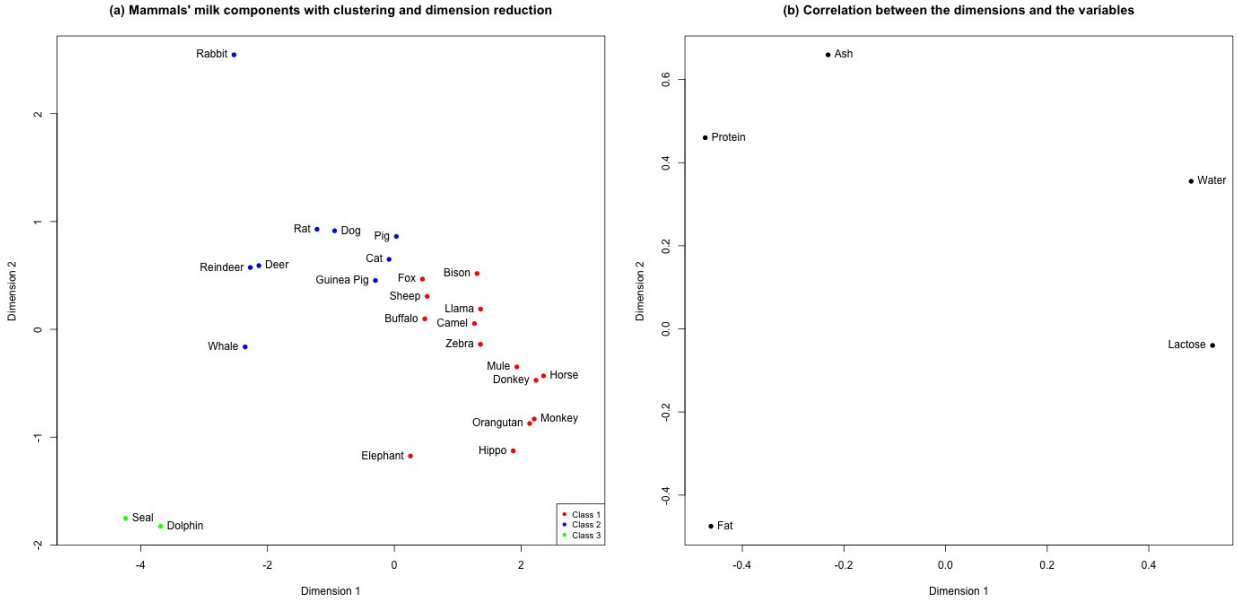

Figure 13: 25 mammals divided in three clusters through CDR is shown in (a). In (b) the milk component variables are projected onto the plane spanned by the two dimensions.

If we look at Figure 13b, we see that all variables reasonably correlate with the dimensions, so there are no variables of no influence. In Figure 13a, it is remarkable that the seals and dolphins are close to each other, whereas the whale lies further from those two animals. This is remarkable, because one would expect that the sea animals have similar milk just like the equines do. The model also clearly captures the similarity between the orangutan and the monkeys, camels and llamas and between a deer and a reindeer. Overall, the CDR model with $\alpha = 0.45$, $c = 3$ and $m = 2$ can make a sensible clustering structure of animals from their milk components.

6 Conclusion

In this thesis a method for the joint selection of parameters for the clustering and dimension reduction (CDR) model is proposed. This model is a generalization of tandem analysis, the FKM method and the RKM method. Within these models, I have shown that the FKM model performs best when it comes to a clustering structure which is masked by randomly generated variables, due to the orthogonality of these variables on the variables that define the clustering structure. Nevertheless, in practice these properties are usually unknown. This thesis shows that, with the usage of the CDR method and thus allowing a mix of these method, the clustering performance is better. This is measured by the Average Silhouette Width (ASW) index and the Calinski-Harabasz pseudo F (pF) index. The parameters that are selected through these indices are the desired number of clusters, number of dimensions and constant α. This is helpful when the researcher has no indication of what these parameters could be in advance. The parameters can be jointly selected, through assessing the clustering performance for different parameter values. This is possible, because the the CDR method is computationally efficient. For future research, I suggest to further assess the performance of the joint selection through the ASW and pF index. This can be done by using a data set where the real values of the parameters are already known. This allows for a comparison between the real clustering structure and the found clustering structure through the proposed methods. Another possible extension would be to apply the joint selection on data sets with categorical, nominal and/or ordinal variables. The CDR method can handle these variables, but I chose not to include this for the sake of a simpler example where the selection methods are exercised.

References

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. *Bagozzi*.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*, 243-256.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*.

De Soete, G., & Carroll, D. J. (1994). K-means clustering in a low-dimensional euclidean space. *New Approaches in Classification and Data Analysis*, 212-219.

Gordon, A. D. (1999). *Classification, 2nd edition*. Chapman & Hall, London.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Markos, A., D'Enza, A. I., & van de Velden, M. (2018). Beyond tandem analysis: Joint dimension reduction and clustering in R. *Journal of statistical software*.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journalof Classification*, 181-204.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edingburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Spector, W. S. (1956). *Handbook of biological data*. Saunders.

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

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

Vichi, M., Vicari, D., & Kiers, H. A. L. (2019). Clustering and dimension reduction for mixed variables. *Behaviormetrika*, *64*, 243–269.

Appendix

Appendix A Tables with data

Table 3: Macroeconomic indicators for 20 OECD countries in 1999.

	GDP	LI	UR	IR	TB	NNS
Australia	1.02	0.53	0.51	0.26	0.04	0.40
Canad	0.68	0.16	0.53	0.25	0.09	0.44
Finland	0.83	-0.06	0.75	0.18	0.50	0.65
France	0.49	0.04	0.74	0.18	0.22	0.62
Spain	0.77	0.16	1.20	0.24	0.07	0.81
Sweden	0.87	0.07	0.56	0.21	0.40	0.34
USA	0.87	0.09	0.28	0.28	-0.08	0.59
Netherlands	0.62	0.10	0.27	0.18	0.40	1.34
Greece	0.68	0.04	0.65	0.58	-0.47	0.68
Mexico	0.49	0.35	0.20	1.04	0.00	1.08
Portugal	0.60	-0.47	0.31	0.24	-0.50	1.19
Austria	0.23	0.04	0.30	0.19	-0.03	0.80
Belgium	0.30	-0.01	0.61	0.18	0.26	1.05
Denmark	0.21	0.09	0.34	0.20	0.19	0.42
Germany	0.17	-0.13	0.60	0.18	0.09	0.65
Italy	0.19	-0.03	0.78	0.30	0.25	0.69
Japan	0.02	0.34	0.27	0.04	0.07	1.28
Norway	0.30	0.06	0.21	0.22	0.41	1.28
Switzerland	0.23	0.13	0.24	0.09	0.25	1.12
UK	0.26	0.31	0.41	0.38	-0.03	0.41

Appendix A Tables with data

Table 4: Milk composition of 25 mammals

	water	protein	fat	lactose	ash
Horse	90.10	2.60	1.00	6.90	0.35
Orangutan	88.50	1.40	3.50	6.00	0.24
Monkey	88.40	2.20	2.70	6.40	0.18
Donkey	90.30	1.70	1.40	6.20	0.40
Hippo	90.40	0.60	4.50	4.40	0.10
Camel	87.70	3.50	3.40	4.80	0.71
Bison	86.90	4.80	1.70	5.70	0.90
Buffalo	82.10	5.90	7.90	4.70	0.78
Guinea Pig	81.90	7.40	7.20	2.70	0.85
Cat	81.60	10.10	6.30	4.40	0.75
Fox	81.60	6.60	5.90	4.90	0.93
Llama	86.50	3.90	3.20	5.60	0.80
Mule	90.00	2.00	1.80	5.50	0.47
Pig	82.80	7.10	5.10	3.70	1.10
Zebra	86.20	3.00	4.80	5.30	0.70
Sheep	82.00	5.60	6.40	4.70	0.91
Dog	76.30	9.30	9.50	3.00	1.20
Elephant	70.70	3.60	17.60	5.60	0.63
Rabbit	71.30	12.30	13.10	1.90	2.30
Rat	72.50	9.20	12.60	3.30	1.40
Deer	65.90	10.40	19.70	2.60	1.40
Reindeer	64.80	10.70	20.30	2.50	1.40
Whale	64.80	11.10	21.20	1.60	0.85
Seal	46.40	9.70	42.00	0.00	0.85
Dolphin	44.90	10.60	34.90	0.90	0.53

Appendix B R code

```
1      install.packages("clustrd")
2    library("clustrd")
3    install.packages("Hmisc")
4    library("Hmisc")
5    #install.packages("fpc")
6    library(fpc)
7    library(cluster)
8
9  # reproduce the results given in the paper of Vichi and Kiers
10
11   setwd("~/Documents/Erasmus Universiteit/Econometrie & Operationele
         Research/Bachelor 3/Thesis/Reproducing")
12
13 set.seed(777)
14
15   # construct the datapoints
16   betweenpoints <- 1
17   betweencentroids <- 6
18   centrex <- c(-1.5*betweenpoints, -betweenpoints, -betweenpoints, -0.5
         *betweenpoints, -0.5*betweenpoints, -0.5*betweenpoints, 0, 0, 0.5*
         betweenpoints, 0.5*betweenpoints, 0.5*betweenpoints, betweenpoints
         , betweenpoints, 1.5*betweenpoints)
19   centrey <- c(0, betweenpoints, -betweenpoints, 2*betweenpoints, 0, -2
         *betweenpoints, betweenpoints, -betweenpoints, 2*betweenpoints, 0,
          -2*betweenpoints, betweenpoints, -betweenpoints, 0)
20  centrey <- centrey - 2
21    iclus <- cbind(centrex, centrey)
22   anglesin <- sin(60*0.0174532925)
23   leftclus <- iclus
24   rightclus <- iclus
25   upclus <- iclus
26
27
28
29
30   rightclus[,1] <- iclus[,1] + betweencentroids
31   upclus[,1] <- iclus[,1] + (betweencentroids/2)
```

```
32  upclus[,2] <- sin(60*0.0174532925)*betweencentroids+ iclus[,2]
33  upclus2 <- upclus
34  upclus2[,1] <- upclus[,1] + betweencentroids
35  #upclus2[,2] <- upclus[,2] + betweencentroids
36
37
38  datapoints <- rbind(leftclus, rightclus, upclus)
39  datapoints2 <- rbind(leftclus, rightclus, upclus, upclus2)
40  #datapoints <- read.csv("datapoints.csv", header = FALSE)
41  maskingpoints <- as.data.frame(matrix(rnorm(42*4, mean = 0, sd = 6),
        ncol = 4))
42  maskingpoints2 <- as.data.frame(matrix(rnorm(56*4, mean = 0, sd = 6),
        ncol = 4))
43  testclusters <- cbind(datapoints, maskingpoints)
44  testclusters2 <- cbind(datapoints2, maskingpoints2)
45
46
47
48  #figure 1
49
50  testfig1 <- kmeans(testclusters, 3)
51  jpeg("fig1.jpg")
52  plot(testclusters$centrex, testclusters$centrey,
53       pch = 16,
54       xlab ="", ylab = "",
55       xlim = c(-2,8), ylim = c(-6,6))
56  text(testclusters$centrex, testclusters$centrey, labels = testfig1$
        cluster, pos = 3)
57  text(0, -5, labels = "cluster 1")
58  text(3, 0.4, labels = "cluster 2")
59  text(6, -5, labels = "cluster 3")
60  dev.off()
61
62
63
64
65  #figure 2
66  initclus <- c
```

```
    (1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3
67  testpca <- prcomp(testclusters, scale = TRUE, center = TRUE)
68  testfig2a <- kmeans(testpca$x[,1:2], 3)
69  testfig2b <- kmeans(testpca$x[,1:3], 3)
70  testfig2c <- kmeans(testpca$x[,1:4], 3)
71  testfig2d <- kmeans(testpca$x[,1:5], 3)
72  jpeg("fig2.jpg", width = 800, height = 800)
73  par(mfrow=c(2,2))
74  plot(-testpca$x[,2], -testpca$x[,1], main="(a) first two principal
        components",
75       xlim = c(-4,3), ylim = c(-6,8),
76       pch = 16,
77       xlab="", ylab = "")
78  text(-testpca$x[,2], -testpca$x[,1], labels = initclus, pos = 3)
79
80  plot(testclusters$centrex, testclusters$centrey, main="(b) first
        three principal components",
81       xlim = c(-2,8), ylim = c(-6,6),
82       pch = 16,
83       xlab="", ylab = "")
84  text(testclusters$centrex, testclusters$centrey, labels = testfig2b$
        cluster, pos = 3)
85
86  plot(testclusters$centrex, testclusters$centrey, main="(c) first four
        principal components",
87       xlim = c(-2,8), ylim = c(-6,6),
88       pch = 16,
89       xlab="", ylab = "")
90  text(testclusters$centrex, testclusters$centrey, labels = testfig2c$
        cluster, pos = 3)
91
92  plot(testclusters$centrex, testclusters$centrey, main="(d) first five
        principal components",
93       xlim = c(-2,8), ylim = c(-6,6),
94       pch = 16,
95       xlab="", ylab = "")
96  text(testclusters$centrex, testclusters$centrey, labels = testfig2d$
```

```
    cluster, pos = 3)
dev.off()
par(mfrow=c(1,1))



# table 1

sink("table1.txt")
print(summary(testpca))
sink()
print(xtable(summary(testpca), type = "latex"), file = "table1.tex")



#figure 3
outrkm <- cluspca(testclusters, 3, 2, seed = 27, scale = FALSE,
    center = FALSE)

jpeg("fig3.jpg", width = 800, height = 800)
plot(outrkm$obscoord[,1], outrkm$obscoord[,2], main="k-means
    clustering in low dimensional space",
      #xlim = c(-2,3), ylim = c(-2,3),
      pch = 16,
      xlab ="Dimension 1", ylab = "Dimension 2")

text(outrkm$obscoord[,1], outrkm$obscoord[,2], labels = initclus, pos
    = 3)
dev.off()



#table 2
outfkm <- cluspca(testclusters, 3, 2, method = "FKM", rotation = "
    varimax", seed = 1, scale = F, center = F)
```

```
130 sink("table2.txt")
131
132 print(outfkm$attcoord)
133 sink()
134 print(xtable(outfkm$attcoord, type = "latex"), file = "table2.tex")
135
136
137
138
139
140 #figure 4
141
142 jpeg("fig4.jpg", width = 800, height = 800)
143 plot(outfkm$obscoord[,1], outfkm$obscoord[,2], main="factorial k-
        means",
144      # xlim = c(-2,2), ylim = c(-2,6),
145       pch = 16,
146       xlab ="Dimension 1", ylab = "Dimension 2")
147
148 text(outfkm$obscoord[,1], outfkm$obscoord[,2], labels = outfkm$
        cluster, pos = 3)
149 dev.off()
150
151 # standardization and correlation
152 data("macro")
153 write.csv(macro,'macrodata.csv', row.names = TRUE)
154 z_macro <- macro
155 correlation <- rcorr(as.matrix(z_macro))
156 sink("correlation.txt")
157 print(correlation)
158 sink()
159
160
161
162
163 #table 3
164
165 macrofkm <- cluspca(macro, 3, 2, method = "FKM", scale = TRUE, center
```

```
    = TRUE, seed = 39, alpha = 0)
z_macro2 <- z_macro
 z_macro$cluster <- macrofkm$cluster
z_macro <- z_macro[order(z_macro$cluster),]
sink("table3.txt")
print(z_macro)
sink()
print(xtable(z_macro, type = "latex"), file = "table3.tex")



#table 4

sink("table4.txt")
print(macrofkm$attcoord)
sink()
print(xtable(macrofkm$attcoord, type = "latex"), file = "table4.tex")

#figure 5
jpeg("fig5.jpg", width = 1400, height = 700)
par(mfrow=c(1,2))
plot(macrofkm$obscoord[,1], macrofkm$obscoord[,2], main="(a) OECD
    with factorial k-means",
      xlim = c(-2.5,2.5), ylim = c(-2.5,2),
      pch = 16,
      col = ifelse(macrofkm$cluster==1,"red",ifelse(macrofkm$cluster
          ==2,"blue","green")),
     xlab = "Dimension 1", ylab = "Dimension 2")
legend("bottomright", pch=16, col = c("red", "blue", "green"), c("
    Class 1", "Class 2", "Class 3"), bty="o", cex=.8)
oecd <- rownames(macrofkm$obscoord)
macrofkm$leftnames <- oecd
macrofkm$rightnames[20] <- ''
macrofkm$rightnames[c(10,13,15,18,19)] <- macrofkm$leftnames[c
    (10,13,15,18,19)]
macrofkm$leftnames[c(10,13,15,18,19)] <- ''
text(macrofkm$obscoord[,1], macrofkm$obscoord[,2], labels = macrofkm$
```

```
    leftnames, pos = 2)
199 text(macrofkm$obscoord[,1], macrofkm$obscoord[,2], labels = macrofkm$
    rightnames, pos = 4)
200
201 plot(macrofkm$attcoord,
202      main = "(b) Correlation between the dimensions and the variables
         ",
203      xlim = c(-0.6,0.7),
204      xlab = "Dimension 1",
205      ylab = "Dimension 2",
206      pch = 16)
207 text(macrofkm$attcoord[,1], macrofkm$attcoord[,2], labels = c("GDP",
    "LI", "UR", "IR", "TB", "NNS"), pos = 4)
208
209 dev.off()
210
211
212 #table 5
213 macropca <- prcomp(macro, scale = TRUE, center = TRUE)
214 sink("table5.txt")
215 print(macropca$rotation[,1:2])
216 sink()
217 print(xtable(macropca$rotation[,1:2], type = "latex"), file = "table5
    .tex")
218
219
220 #figure 6
221 jpeg("fig6.jpg", width = 800, height = 800)
222 macrotkm <- kmeans(macropca$x[,1:2], 3)
223
224 plot(-macropca$x[,2], -macropca$x[,1], main="OECD with tandem
    analysis",
225      xlim = c(-4,2), ylim = c(-2.5,2.5),
226      pch = ifelse(macrotkm$cluster==1,23,ifelse(macrotkm$cluster
         ==2,22,24)),
227      xlab ="", ylab = "")
228 legend("bottomright", pch=c(23,22,24), c("Class 1", "Class 2", "Class
     3"), bty="o", cex=.8)
```

```
229 macrotkm$leftnames <- oecd
230 macrotkm$rightnames[20] <- ''
231 macrotkm$rightnames[c(4,6,8,10,13,15,19)] <- macrotkm$leftnames[c
    (4,6,8,10,13,15,19)]
232 macrotkm$leftnames[c(4,6,8,10,13,15,19)] <- ''
233
234
235
236 text(-macropca$x[,2], -macropca$x[,1], labels = macrotkm$leftnames,
    pos = 2)
237 text(-macropca$x[,2], -macropca$x[,1], labels = macrotkm$rightnames,
    pos = 3)
238 dev.off()
239
240 #example with 4 simulated clusters
241 jpeg("fig4clus.jpg", width = 800, height = 400)
242 par(mfrow=c(1,2))
243 outfkm2 <- cluspca(testclusters2, 3, 2, method = "FKM", rotation = "
    varimax", seed = 1, scale = F, center = F)
244 plot(testclusters2$centrex, testclusters2$centrey, main="(a)
    factorial k-means, c = 3",
245       xlim = c(-2,11), ylim = c(-4,7),
246       pch = 16,
247      xlab ="", ylab = "")
248
249   text(testclusters2$centrex, testclusters2$centrey, labels = outfkm2
      $cluster, pos = 3)
250
251
252   outfkm3 <- cluspca(testclusters2, 4, 2, method = "FKM", rotation =
      "varimax", seed = 1, scale = F, center = F)
253   plot(testclusters2$centrex, testclusters2$centrey, main="(b)
      factorial k-means, c = 4",
254     xlim = c(-2,11), ylim = c(-4,7),
255     pch = 16,
256     xlab ="", ylab = "")
257   text(testclusters2$centrex, testclusters2$centrey, labels = outfkm3
      $cluster, pos = 3)
```

```
258 dev.off()
259
260 jpeg("4clusindex.jpg", width = 800, height = 400)
261 par(mfrow=c(1,2))
262 numclus <- c(3:10)
263 pf = tuneclus(testclusters2, 3:10, 2:2, method = "FKM", criterion
      = "ch", dst = "full")
264 plotpf <- cbind(numclus, as.numeric(data.matrix(pf$critgrid$X2)))
265 plot(plotpf, main = "Calinski-Harabasz pseudo F index", type = "o",
      pch = 16, xlab = "number of clusters", yaxt = "n")
266
267 asw = tuneclus(testclusters2, 3:10, 2:2, method = "FKM",
    criterion = "asw", dst = "full")
268 plotasw <- cbind(numclus, as.numeric(data.matrix(asw$critgrid$X2)))
269 plot(plotasw, main = "Average Silhouette Width index", type = "o",
    pch = 16, xlab = "number of clusters", yaxt = "n")
270
271
272
273
274 dev.off()
275 par(mfrow=c(1,1))
276
277
278 # parameter selection for OECD 1999
279 jpeg("oecdparam.jpg", width = 800, height = 400)
280 par(mfrow=c(1,2))
281
282 oecdpf = tuneclus(macro, 3:10, 2:3, method = "FKM", criterion = "
    ch", dst = "full")
283 plotoecdpf <- cbind(numclus, as.numeric(data.matrix(oecdpf$critgrid
    $X2)), as.numeric(data.matrix(oecdpf$critgrid$X3)))
284 plot(cbind(numclus, plotoecdpf[,2]), type = "o", pch = 18, col = "
    blue", xlab = "number of clusters", main = "Calinski-Harabasz
    pseudo F index", yaxt = "n")
285 lines(cbind(numclus, plotoecdpf[,3]), type = "o", pch = 18, col = "
    black")
286 text(10, plotoecdpf[8,2], "m = 2", pos = 2, col = "blue")
```

```
287     text(10, plotoecdpf[8,3], "m = 3", pos = 2, col = "black")
288
289
290     oecdasw = tuneclus(macro, 3:10, 2:3, method = "FKM",    criterion =
            "asw", dst = "full")
291     plotoecdasw <- cbind(numclus, as.numeric(data.matrix(oecdasw$
            critgrid$X2)), as.numeric(data.matrix(oecdasw$critgrid$X3)))
292     plot(cbind(numclus, plotoecdasw[,2]), type = "o", pch = 18, col = "
            blue", xlab = "number of clusters", main = "Average Silhouette
            Width index", yaxt = "n")
293     lines(cbind(numclus, plotoecdasw[,3]), type = "o", pch = 18, col =
            "black")
294     text(10, plotoecdasw[8,2], "m = 2", pos = 2, col = "blue")
295     text(10, plotoecdasw[8,3], "m = 3", pos = 2, col = "black")
296
297
298     dev.off()
299     par(mfrow=c(1,1))
300
301
302     # variable alpha
303     aswval <- matrix(0, nrow = 20, ncol = 1)
304       pfval <- matrix(0, nrow = 20, ncol = 1)
305     for (p in seq(from = 0.05, to = 1, by = 0.05)) {
306       tuneclusalphapf <- tuneclus(macro, nclusrange = 3, ndimrange = 2,
              alpha = p, criterion = "ch")
307       tuneclusalphaasw <- tuneclus(macro, nclusrange = 3, ndimrange =
              2, alpha = p, criterion = "ch")
308       val <- 20*p
309       pfval[val,1] <- tuneclusalphapf$critbest
310       aswval[val,1] <- tuneclusalphaasw$critbest
311
312     }
313
314     jpeg("oecdalpha.jpg", width = 800, height = 400)
315     par(mfrow=c(1,2))
316     chval2 <- rbind(43, chval)
317     aswval2 <- rbind(0.094, aswval)
```

```
318     plot(cbind(seq(from = 0.00, to = 1, by = 0.05), chval2), type = "o"
            , pch = 16,
319           xlab = "  ", ylab = "", main = "Calinski-Harabasz pseudo F
                 index", yaxt = "n")
320     plot(cbind(seq(from = 0.00, to = 1, by = 0.05), aswval2), type = "o
            ", pch = 16,
321           xlab = "  ", ylab = "", main = "Average Silhouette Width index
                 ", yaxt = "n")
322     dev.off()
323     par(mfrow=c(1,1))
324
325
326     #CDR macro data alpha = 0.2 c = 3 m = 2
327
328     oecdcdr <- cluspca(macro, 3, 2, alpha = 0.2, center = T, scale = T,
            seed = 31101998)
329
330     jpeg("oecdcdr.jpg", width = 1400, height = 700)
331     par(mfrow=c(1,2))
332     plot(oecdcdr$obscoord[,1], oecdcdr$obscoord[,2], main="(a) OECD
            with clustering and dimension reduction",
333           #xlim = c(-2.5,2.5), ylim = c(-2.5,2),
334           pch = 16,
335           col = ifelse(oecdcdr$cluster==1,"red",ifelse(oecdcdr$cluster
                 ==2,"blue","green")),
336           xlab = "Dimension 1", ylab = "Dimension 2")
337     legend("bottomright", pch=16, col = c("red","blue","green"), c("
            Class 1", "Class 2", "Class 3"), bty="o", cex=.8)
338     oecdcdr2 <- rownames(oecdcdr$obscoord)
339     oecdcdr$leftnames <- oecdcdr2
340     oecdcdr$rightnames[20] <- ''
341     oecdcdr$rightnames[c(3,9,13,16)] <- oecdcdr$leftnames[c(3,9,13,16)]
342     oecdcdr$leftnames[c(3,9,13,16)] <- ''
343     text(oecdcdr$obscoord[,1], oecdcdr$obscoord[,2], labels = oecdcdr$
            leftnames, pos = 2)
344     text(oecdcdr$obscoord[,1], oecdcdr$obscoord[,2], labels = oecdcdr$
            rightnames, pos = 4)
345
```

```
346     plot(oecdcdr$attcoord,
347          main = "(b) Correlation between the dimensions and the
                 variables",
348         #xlim = c(-0.6,0.7),
349          xlab = "Dimension 1",
350          ylab = "Dimension 2",
351          pch = 16)
352     text(oecdcdr$attcoord[,1], oecdcdr$attcoord[,2], labels = c("GDP",
        "LI", "UR", "IR", "TB", "NNS"), pos = 4)
353
354     dev.off()
355
356
357     # mammals
358     help <- rep(0, 20*2*8)
359     mammalsindex <- array(help, c(8,2,20))
360
361     for (q in seq(from = 0.05, to = 1, by = 0.05)) {
362     milkasw <- tuneclus(mammals, nclusrange = 3:10, ndimrange = 2:3,
        alpha = q, criterion = "asw", method = "FKM")
363     vallie <- 20*q
364     mammalsindex[,1,vallie] <- as.numeric(data.matrix(milkasw$critgrid$
        X2))
365     mammalsindex[,2,vallie] <- as.numeric(data.matrix(milkasw$critgrid$
        X3))
366
367
368     }
369     milkclus <- cluspca(mammals, 3, 2, alpha = 0.45, seed = 1)
370
371     jpeg("milkclus.jpg", width = 1400, height = 700)
372     par(mfrow=c(1,2))
373     plot(milkclus$obscoord[,1], milkclus$obscoord[,2], main="(a)
        Mammals' milk components with clustering and dimension reduction
        ",
374          xlim = c(-5,3),
375          #ylim = c(-2.5,2),
376          pch = 16,
```

```
377         col = ifelse(milkclus$cluster==1,"red",ifelse(milkclus$cluster
                ==2,"blue","green")),
378         xlab = "Dimension 1", ylab = "Dimension 2")
379     legend("bottomright", pch=16, col = c("red", "blue", "green"), c("
            Class 1", "Class 2", "Class 3"), bty="o", cex=.8)
380     milkclus2 <- rownames(milkclus$obscoord)
381     milkclus$leftnames <- milkclus2
382     milkclus$rightnames <- ''
383     milkclus$rightnames[c(1,3,17,21,24,25)] <- milkclus$leftnames[c
            (1,3,17,21,24,25)]
384     milkclus$leftnames[c(1,3,17,21,24,25)] <- ''
385     text(milkclus$obscoord[,1], milkclus$obscoord[,2], labels =
            milkclus$leftnames, pos = 2)
386     text(milkclus$obscoord[,1], milkclus$obscoord[,2], labels =
            milkclus$rightnames, pos = 4)
387
388     plot(milkclus$attcoord,
389         main = "(b) Correlation between the dimensions and the
                variables",
390         #xlim = c(-0.6,0.7),
391         xlab = "Dimension 1",
392         ylab = "Dimension 2",
393         pch = 16)
394     text(milkclus$attcoord[,1], milkclus$attcoord[,2], labels = c("
            Water", "Protein", "Fat", "", "Ash"), pos = 4)
395     text(milkclus$attcoord[,1], milkclus$attcoord[,2], labels = c("", "
            ", "", "Lactose", ""), pos = 2)
396
397
398     dev.off()
```