Erasmus University Rotterdam

Erasmus School of Economics

MASTER THESIS - ECONOMETRICS AND MANAGEMENT SCIENCE

BUSINESS ANALYTICS AND QUANTITATIVE MARKETING

Clustering and Dimension Reduction of Rating Data

A Comparison Study of Different Clustering and Dimension Reduction Methods Applied on Rating Data

Author:
J.C. Borges Soares
Student ID number:
473966

Supervisor:
M. van de Velden
Second assessor:
A. Archimbaud

April 23, 2023

Abstract

In this research, a comparison study of various clustering and dimension reduction methods that were applied to diverse types of rating data coding is performed. A simulation study is conducted to investigate the efficacy of these methods by controlling different parameters to identify the most optimal performing method under distinct circumstances. The clustering methods were classified into three categories, namely, full-dimensional clustering, successive dimension reduction and clustering, and simultaneous dimension reduction and clustering. The data coding was categorised into three types, namely, raw data, doubled data, and categorical data. The findings revealed that, in the presence of noise, the best performing method was a simultaneous dimension reduction and clustering approach. This method is called doubled reduced K-means. However, when the data set lacked noise, the full-dimensional clustering method, K-means, emerged as the most effective approach. The research is concluded by implementing the reduced K-means method, which was found to be the best-performing approach, on a real-world data set to showcase the visualisation of the results.

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Contents

1 Introduction

In the marketing field, the utilisation of customer data is becoming more and more popular, because it provides valuable insights into customer behaviour, which can be of great value to companies. One of the most commonly used types of data in this field is rating data, which is obtained by having customers rate products or services on a specified point scale. The use of rating data enables companies to personalise their advertisements by comparing customers with similar preferences to each other, potentially leading to increased revenue. For example, consider a customer who frequently orders takeaway food and is prompted to rate their satisfaction on a scale of 1 to 5 after each delivery. By collecting and analysing this data, a preference profile can be constructed for each customer, highlighting the types of food they prefer. By comparing preference profiles across different customers, it is hereby made possible to identify which customers have similar preferences, allowing for more targeted advertising. For instance, if one customer with a certain preference profile rates a new type of food highly, it is likely that other customers with similar preference profiles also enjoy that type of food.

As the number of customers increases, the difficulty of directly comparing all customers goes up. One method to address this challenge is by utilising clustering methods, which group observations based on their similarity across variables. This approach allows for the identification of groups of customers with similar preferences and distinct preferences relative to other groups. When a new customer is introduced, they can be compared to the various customer clusters and targeted with products or services preferred by the group they are most similar to.

The ever-growing use of data also means that data sets are growing in size and complexity, with both the number of observations and variables increasing significantly. This presents challenges when applying clustering methods to high-dimensional data. To address this issue, reducing the number of variables in the data set is often necessary. One approach could be to manually select a subset of variables that are deemed representative and discard the remaining variables. However, this can limit the usefulness of the data, as all observations may contain useful information. Thus, there is need for methods that can improve the interpretability of large, high-dimensional data sets while still preserving the relevant information.

Principal component analysis (PCA) is a method used to increase the interpretability of high-dimensional data sets. PCA aims to reduce the number of dimensions, or variables. This results in a smaller number of new variables being created, which are a linear combinations of the original variables. The goal here is to maximise the amount of information retained in the data while minimising the loss of information due to dimensionality reduction. This approach can improve the interpretability of the data while preserving the key information contained in the original data set.

The combination of PCA and the previously discussed clustering methods can aid in discovering structure within complex and difficult to interpret high-dimensional data sets. One approach is to apply these methods sequentially, with PCA performed prior to a cluster analysis. However, alternative methods have been developed that combine the two approaches in a single approach, such as reduced or factorial K-means. This is because the sequential application of PCA and cluster analysis optimises two distinct objective function, which can result in sub-optimal outcomes in certain cases. By integrating the objectives of both methods, a more balanced and effective approach is achieved.

The suitability of a particular clustering analysis given a data set is not solely determined by the dimensionality of the data, but it is also influenced by other characteristics of the data, such as: the variance of the variables, the correlation between variables, the presence of noise, and the degree of overlap in the rating distribution among different variables. Consequently, selecting an appropriate clustering method for a given data set can be a complex task. In this research, we aim to evaluate and compare the performance of various clustering methods on rating data and identify the method that is most effective in specific situations. The research question guiding this research is as follows:

> *How do different methods of cluster analyses combined with dimension reduction perform when applied on rating data?*

The remainder of this thesis is structured as follows: In Section 2, we discuss previously done research on this topic. In Section 3, we outline the various methods employed to conduct this research. Section 4 details the simulation study that was performed to compare the performance of the different methods. Section 5 provides an illustrative example of the different methods applied to a real-world dataset. Finally, in Section 6, we present a summary of our findings and offer suggestions for future research in this area.

2 Literature Review

A significant amount of research has been devoted to exploring techniques that cluster rating data sets and reduce the number of dimensions of these data sets. In this section, a review and comparison of existing studies on this topic is presented.

2.1 Coding of rating data

As discussed in the Introduction, there are many useful methods that can help interpreter and summarise large data sets. This research focuses mainly on rating data. However, this does not mean that all the analyses are performed on the raw rating data. Researches have introduced many methods for coding this data, and some are discussed here.

2.1.1 Rating data

Rating data is a type of data that enables respondents to give a rating value to a certain statement. It allows the respondents to show preferences. Harpe (2015) discusses one of the main issues raised when analysing rating scale data: are rating scale data ordinal or interval? This question arises when looking at the nature of rating data. One criticism is that the distance between two subsequent categories is assumed to be equal. If the rating data is interval, this property would be automatically assumed due to the nature of an interval variable. Ordinal variables do not assume a distance between variables, but only the fact that one variable has a "higher" ranking than the other. Thus, ordinal variables do not assume that the distance between the categories are similar. Why this difference is of importance, is because not all statistical analyses that are used on interval data can be used on ordinal data. A simple example is the calculation of the mean statistic, which is a statistic used to show the "middle" of a set of data. With interval data, the mean can be calculated by taking the average value of a variable, which is the sum of all variables divided by the total number of variables. When considering ordinal data, calculating the mean statistic is not appropriate, because one does not assume that the distance between categories is constant. One should use, for example, the median statistic to show the middle of the data set. The research of Harpe (2015) gives some recommendations regarding the use of certain statistical analyses. One of the most considerable recommendations is the fact that rating values with a numerical value and a point-scale of at least five, may be treated as interval data. Hsu & Feldt (1969) showed that with a minimum of five categories, the use of analyses used for interval data are appropriate. When conducting our research, the importance of acknowledging the differences between ordinal and interval rating data should be noted.

2.1.2 Paired comparison and rank order data

Another type of data used to indicate preferences of individuals are paired comparison and rank order data. Paired comparison data refers to data that compares two observations to decide which of these observations is preferred (Guttman, 1946). Rank order data is a type of data that is obtained by ranking a larger group of observations. The two methods are closely related, as both methods give a ranking order of ones preferences. However, there are two notable limitations differentiating these two types of data. The first limitation is that paired comparison data allows for inconsistency in preferences, as paired comparison data compares two observations at a time, which could lead to situations where, for example, product A is favoured over product B, product B is picked over product C, but product C is preferred over product A. This is called a transitive relation and this relation can not be shared when using paired comparison data, as opposed to rank order data. This would imply that rank order data is preferable, however, the second limitation is that it can be more difficult for participants to rank a large set of observations at the same time, compared to a set of two observations at a time, which would make paired comparison data more preferable.

One drawback for both of the two types of data is that they only allow participants to give a ranking of the data relative to other observation, thus, not allowing a participant to show preference or dislike to all objects. This should be taken into account when dealing with these types of data.

2.1.3 Successive categories data

Another type of ordinal data, is the so-called successive categories data (Torgerson, 1958). Successive categories data is a type of categorical data which has a ranking for each category. This means that the data does have a natural ranking, but does not have to be numerical. One example for this type of data is educational level. Where it is clear that there is a hierarchy in the different educational levels, but there is no numerical value attached to each value.

2.1.4 Doubled matrix for rank rating data

Greenacre (1984) discusses an important point about rating data, which is often overlooked. This is the property that rating data should be invariant to the "direction" of the associated statements. For example, if the statement "The present government has a sound economic policy" is given a rating of 4 on a scale of 5, the statement "The present government does not have a sound economic policy" should result in a rating of 1 on a scale of 5. In other words, a rating of agreement is equivalent to a complementary rating of disagreement. To take the absolute and bipolar nature of the rating scale data into account, Greenacre (1984) introduces the concept of "doubling" rating data. Here, each variable is reflected with respect to the point scale. Which means that if the original variable has a value of 5 on a 7-point rating scale, a new variable is created which has a value of 2 on the reversed 7-point scale. In other words, a variable is added that answers the complementary statement. Applying correspondence analysis on this doubled rating matrix leads to different results compared to correspondence analysis on the raw data set. Greenacre shows that, when applying correspondence analysis on this doubled matrix, the points in the biplot become more spread out compared to the results when applied solely on the raw rating data. This is the case because analysing solely the raw rating data, only takes positive association into account. The impact of doubling the rating data is not as large when the answers of different clusters or individuals are spread out. However, when having situations where different clusters tend to have similar preferences, we see that the use of doubled rating data can be of relevance.

2.2 Cluster analyses and dimension reduction

Now that we have discussed different methods of coding rating scale data, we take a look at different analyses applicable to these different types of data. Here we take a look at cluster analyses, which focuses on clustering individuals in groups with similar preferences. In this research, the clustering methods can be divided in three groups: clustering of the raw data set, successively applying dimension reduction and clustering analysis and simultaneously reducing the number of dimensions and clustering the data.

2.2.1 K-means

MacQueen (1967) introduced a clustering algorithm which is widely known for its simple use and easy application. The so-called K-means algorithm finds a cluster allocation for which the distance between observations in the same cluster is as small as possible, whilst the distance between observations in a different cluster is maximised. The algorithm is initiated by selecting k (random) cluster centroids. Every iteration, each observation is allocated to the closest centroid. Thereafter, the centroids are recalculated as the centre of all observation in its cluster. This procedure is repeated until the centroids have stabilised or a pre-defined number of iterations is reached. The strength of the K-means algorithm is its relatively good performance and its simplicity. However, it does have weaknesses which needs to be addressed.

The K-means algorithm is sensitive to outliers, for the reason that the centroids are calculated as a mean of different observations (Kaushik & Mathur, 2014). As known, the mean statistic is sensitive to outliers and this results in outliers having a large effect on the K-means algorithm as well. K-means also does tend to perform worse when a data set is of a high dimensionality, meaning that the number of variables is large. A larger number of dimensions can lead to the curse of dimensionality (Bellman & Kalaba, 1959). The curse of dimensionality is a phenomena that occurs when analysing data sets with a high dimensionality. It makes it so, that the number of random samples needed to estimate the objective function grows exponentially with the number of dimensions (Chen, 2009). Another weakness of the K-means algorithm is the fact that the clustering algorithm has trouble clustering data when the underlying clustering structure has varying sizes and densities.

2.2.2 Partitioning around medoids

The K-means algorithm, however, is not the correct method to use when clustering categorical data (Madhulatha, 2011). The K-means algorithm uses the Euclidean distance, which is not defined for categorical variables. Therefore, the k-medoid analysis can be used, specifically the partitioning around medoids (PAM) algorithm (Kaufman & Rousseeuw, 1990). The main difference between K-medoids and K-means is that K-medoids uses dissimilarities instead of the (Euclidean) distance measure. The total sum of dissimilarities results in an objective function called the total deviation, which is optimised when solving the clustering problem. There are different algorithms to optimise the objective function. PAM is one of those used algorithms. The PAM algorithm consists of two phases. These phases are the so-called build phase and swap phase. In the build phase, the initial clustering of the medoids is chosen. In the swap phase, the clustering is improved towards a (local) optimum. The K-medoids approach most likely encounters the same problems that K-means does, this is, it has trouble dealing with high-dimensional data sets as well as unbalanced data.

2.2.3 Principal component analysis

As the above two methods struggle with clustering high-dimensional data sets, methods have been introduced to reduce the number of dimensions of a given data set. One of these methods is principal component analysis (PCA), introduced by Pearson (1901). PCA aims to reduce the number of variables (dimensionality) of a data set whilst keeping as much variability as possible. This is done by creating new variables, which are a linear combination of the original variables. These newly created variables are uncorrelated with each other. These new variables are called principal components (PC's) and finding these PC's is done by solving an eigenvector problem (Jolliffe & Cadima, 2016). Solving the eigenvector problem of the covariance matrix gives pairs of eigenvectors and eigenvalues. The eigenvectors give the directions of the axes that explain the most variance, with the corresponding eigenvalues indicating how much variance this axis explains. Thus, the eigenvectors represents the principal components and the eigenvalues represent the variance explained by each component.

The principal components can be used to apply cluster analysis on, such as K-means, resulting in a so-called tandem analysis. Using this type of cluster analysis, in theory, yields better results when dealing with high-dimensional data sets, as it does not suffer from the curse of dimensionality.

2.2.4 Correspondence analysis and multiple correspondence analysis

PCA assumes interval-scaling of the data, however, when dealing with categorical data, other methods need to be used. One way of reducing the number of dimension of categorical data is correspondence analysis (CA), introduced by Fisher (1940). CA is a method used to display rows and columns of a data set as points in a vector space of a lower dimensionality (Greenacre, 1984). CA is similar to PCA, however, CA is optimised such that it is applicable to categorical data. The way this is done is by using a two-way contingency table. Two-way contingency tables give the count of different choices made when comparing two categorical variable. It can also be applied on rating scale data. Therefore, the ratings need to be converted to categories. This can be done by making a category of each rating value, or by creating groups of rating values. In the two-way contingency table, the rows are given by one categorical variable, and the columns by another. Each cell then gives the number of occurrences where a combination of that row and column is given.

By solving an eigenvalue problem using singular value decomposition (SVD), CA creates components which explain as large of a proportion of variance as possible. Similar to PCA, we can use CA to apply a tandem-analysis on the data, giving us the ability to better cluster the data even with a large number of (noise) variables.

CA can be expanded to multiple correspondence analysis (MCA), which is able to be applied to more than two categorical variables. The way this is done, is by creating an indicator matrix and applying CA on this matrix (Greenacre, 1984).

2.2.5 Optimal scaling of paired comparison data

The research of Guttman (1946) regarding rank order data, discussed in Section 2.1.2, got extended by Van De Velden (2004), who introduced an optimal scaling method for paired comparison data. The method estimates the underlying latent variables that explain the preferences of the respondents and assigns these estimates to observations such that the between-observation variance is as large as possible relative to the overall variance. It is shown that that the optimal scaling of paired comparison data is closely related to the correspondence analysis of paired comparison data.

2.2.6 Tandem analysis

PCA and CA can be used to apply tandem-analysis on the data, resulting in a clustering in a lower dimensional space. This is done by applying a clustering algorithm (such as K-means) on the received components of the PCA or CA methods. The tandem analysis can also be applied in combination with MCA. Here, MCA is applied on the categorical data set and sequentially K-means analysis is applied on the resulting components.

As the dimension reduction methods reduce the number of dimensions, without taking the original cluster structure into account, it could be possible that the reduced dimensions are a combination of original variables that are not much linked to the clustering structure. To combat this problem, certain researchers came up with methods that do not successively reduce the number of dimensions and cluster the data (tandem analysis), however, the number of dimensions are reduced simultaneously with the clustering of the data.

2.2.7 Reduced and factorial K-means

The reduced K-means (RKM) and factorial K-means (FKM) methods, proposed by De Soete & Carroll (1994) and Vichi & Kiers (2001) respectively, are methods that focus on clustering the data whilst reducing the number of dimensions. This is done by optimising an objective function which decomposes the data in such a manner that it closely resembles the original data set, but also reduces the number of dimensions and clusters this data. The objective function is also where the difference between RKM and FKM lies. The RKM model tries to fit the data with the K-means centroids lying in the reduced space. This is done by minimising the total squared distance between observations and centroids in the reduced space. FKM fits the centroids in the reduced space by minimising the within variance, which is the variance of the centroids and the observations, both in the reduced space (Timmerman et al., 2010).

A conclusion drawn by Timmerman et al. (2010) is that for both methods the performance of recovering the cluster membership decreases when the overlap between clusters increases, which is not surprising and generally occurs with every clustering analysis used.

2.2.8 Cluster correspondence analysis

The two methods above are both designed to be applied on interval data. A simultaneous clustering and dimension reduction method that is applicable on categorical data is the cluster correspondence analysis (CCA), introduced by Van De Velden et al. (2017). This method applies CA as well as clusters the data. It is done by use of a "super indicator matrix", which is an matrix containing binary variables that indicate which value is given to a certain statement. This means that this matrix has a column for every possible combination of category value for each variable. CCA is essentially regular CA on the contingency tables of clustering and categorical variables. The data set has the clusters represented on the rows. The obtained row coordinates maximise the between cluster variance. Similarly to the other methods that simultaneously cluster data as well as reduce the number of dimensions, CCA uses a single objective function. This optimisation is done by use of an algorithm. This algorithm improves the solution of the objective function with each iteration until convergence. CCA can be applied in such a way that it results in a biplot which depicts the means of the clusters and variables in a reduced space.

3 Methodology

This research focuses on comparing the performance of different approaches for clustering rating data. This rating data can be coded in several different manners. The different options we chose in this research are: raw rating data, doubled rating data and categorical data. Furthermore, we take a look at different clustering methods. The methods we consider at are: K-means, tandem analysis, reduced K-means, factorial K-means, partitioning around medoids and cluster correspondence analysis.

3.1 Data coding

For the coding of the data, we consider the raw rating data set and the doubled rating data set, which is further explained in Section 3.1.1 and 3.1.2 respectively (Greenacre, 1984). In addition, we are treating the ratings as categories and making dummy variables of all the different ratings and questions, which is described in Section 3.1.3.

3.1.1 Raw rating matrix

The first type of data coding is leaving the rating data as it is, which we call the raw data set. Rating data is ordinal data obtained on a point-scale. This point-scale gives the maximum rating of a certain variable. Thus, having a 5-point-scale means that the ratings range from one to five. In this research, we apply methods that are used for nominal, ordinal and interval data. For interval data, the assumption is made is that the difference between each category is equal (Myers & Winters, 2002). This is not always a good assumption to make, however, Harpe (2015) states in his research that with a sufficient amount of categories, it is not a bad assumption to treat the rating scale data as interval.

3.1.2 Doubled rating matrix

Greenacre (1984) shows a method of analysing rating scale data. Instead of applying analyses on the raw rating data, the data matrix is “doubled”. We use bipolar rating variables, which are ordinal variables that have a lower and higher extreme, or “poles”. One important mathematical property of rating scale data is that the given ratings should not depend on the direction the questions or statements are stated in. This means that when a statement is stated as, for example: “Product X has a nice packaging”, then it should not make a difference when the statement is stated as: “Product X does not have a nice packaging” (apart from the fact that the rating should be the other way around). To take this property into account, a doubled data matrix is used. Every rating value is doubled by its “complement” value. This complement value is the rating value measured from the other end of the point scale spectrum. So if we have a value of 2 on a 11-point scale, the complement value equals 9. The doubled rating data forms a symmetry between the two extreme values of the bipolar variables, which can further be shown when applying correspondence analysis or principal component analysis on this doubled matrix. As these analyses are constant with regarding the scale direction. This is elaborated on in Section 3.2.5.

3.1.3 Indicator matrix

The last method of coding data for the analyses is by handling the data as categorical data. In most analyses, rating data is considered as ordinal data. However, the discrete nature of the data makes it so that we can also treat each rating as a separate category.

The type of categorical data coding used in this research is by means of the “super indicator matrix” $\mathbf{Z}$ (Greenacre, 1984). Let $\mathbf{Z}_j$ be an $N \times q_j$ indicator matrix, with N the number of observations and q_j the length of the rating point scale of variable j. In this matrix, the rows represent the observations and the columns represents the categories. The chosen value for category j is represented as a one in the corresponding column and the remaining elements are zero. This means that each row of $\mathbf{Z}_j$ sums up to one. The super indicator matrix $\mathbf{Z}$ is defined as $\mathbf{Z} = [\mathbf{Z}_1, ..., \mathbf{Z}_Q]$, with Q being the number of variables. $\mathbf{Z}$ has a dimensionality of $N \times T$, where $T = \sum_{j=1}^{Q} q_j$. Thus, the indicator matrix $\mathbf{Z}$ is made up by of the combination of all possible values of all possible variables.

Handling the rating data as categorical data does make it that we do not make the assumption of having an equal spacing between the different rating scale values. Relaxing this assumption can be advantageous in some cases. However, with the assumption of the data being categorical, we do lose the property of ratings having an order. In other words, we can not know if a rating is higher than a certain other rating.

3.2 Cluster analyses

Previously we discussed different clustering methods, which can be divided in multiple categories. The first category applies clustering analyses on the raw data set. The second category first applies dimension reduction methods on the data and thereafter clusters the data. The third and final category uses a simultaneous method of clustering the data whilst reducing the number of dimensions to get a final cluster allocation.

3.2.1 K-means algorithm

The K-means algorithm has first been introduced by MacQueen (1967). It is a clustering algorithm well known for its simple application. The method classifies the observations of the data set in one of K clusters, with K being predefined.

Let $\mathbf{x}_{ik}$ be the vector of the i'th observation from the set of observations in cluster k, C_k. We then get the objective function, which is equal to the total sum of squared error,

$$E = \sum_{k=1}^{K} \sum_{\mathbf{x}_{ik} \in C_k} \|\mathbf{x}_{ik} - \boldsymbol{\mu}_k\|^2. \tag{1}$$

Here $\boldsymbol{\mu}_k$ is the cluster mean of cluster C_k and the distance measure ($\|.\|$) used is the Euclidean distance. That is, if $\mathbf{a} = [a_1, a_2, ..., a_N]$ and $\mathbf{b} = [b_1, b_2, ..., b_N]$ are N-dimensional observations, then the Euclidean distance between these two vectors can be obtained as follows:

$$\|\mathbf{a} - \mathbf{b}\| = \left[\sum_{i=1}^{N} (a_i - b_i)^2\right]^{1/2}. \tag{2}$$

To minimise the objective function in Equation (1), an algorithm is used. This algorithm consists of three steps. The first step is to initiate the clustering allocation. This can be done by randomly assigning each observation to a cluster, or by allocating the observations based on previous findings. Given a cluster allocation, the second step is to calculate the mean of each cluster, by taking the mean of all observations in that cluster. This gives K cluster means, hence the name K-means. Given the means, the third step is to reallocate the observations into the cluster to which mean it is closest to. With the newly allocated clusters, the means can be recalculated and this algorithm can be iteratively continued until the objective function is minimised. If the objective function does not stabilise, the algorithm can be aborted by setting a maximum number of iterations.

Using the algorithm stated above, the objective function always decreases with each step until convergence. This because each iteration, the cluster allocation is updated by allocating the observations to a cluster to which mean it is closest to compared to the iteration before, making it impossible to increase the distance from an observation to a mean, and thus, making it impossible to increase the objective value. The problem that may arise with this approach, is that the algorithm ends up in a local minimum. Because the objective value always decreases, when a certain optimum is found, it is impossible to get to another optimum if this makes the objective function increase in the following iterations. There are methods to combat this problem. The easiest to implement and most intuitive method is to start with more random starting values and choosing the cluster allocation that leads to the lowest objective value. Using more starting values increases the probability of the algorithm finding the global minimum. This method is applied in this research.

One of the major drawbacks of the K-means algorithm is the so called curse of dimensionality, introduced by Bellman & Kalaba (1959), which occurs when the data is of high dimensionality. When clustering data, adding variables to the data set increases the volume of the data set exponentially, with the volume being the possible samples that can be drawn from a data set. Adding variables to the data set increases the distance between observations, because the distance is calculated over all variables. This can lead to the problem that when observations that are supposed to be in the same cluster, are allocated to different clusters when variables are added, because of the increased distance between these observations. This is especially the case when variables that are added are not relevant to the underlying clustering structure. Variables that are not relevant to the clustering structure are called "noise" variables. These variables are not relevant to the clustering structure, but are taken into account when applying full-data clustering (Vichi & Kiers, 2001). The curse of dimensionality can lead to lower clustering accuracy.

3.2.2 Partitioning around medoids

When using the indicator matrix categorical coding, it is not preferred to use the K-means algorithm for clustering the data. The K-means algorithm uses the Euclidean distance measure to define the distance between variables. However, when coding the data using the indicator matrix, it is coded such that the data is categorical, for which the Euclidean distance is not defined. Thus, we need to use an algorithm which is applicable to categorical data. The method we consider in this research is K-medoids, specifically the so-called partitioning around medoids (PAM) algorithm (Kaufman & Rousseeuw, 1990).

Instead of calculating the mean of observations in each cluster, as one would normally do in the K-means algorithm, the PAM algorithm finds the medoids of every cluster. The medoids are calculated by finding an observation $\mathbf{x}_m$ in each cluster that minimises $\sum_{\mathbf{x}_n \in C_m} \text{dis}(\mathbf{x}_m, \mathbf{x}_n)$, where C_m is the cluster containing observation $\mathbf{x}_m$, $m = 1, ..., N$, $n = 1, ..., N$ and $\text{dis}(\mathbf{x}_m, \mathbf{x}_n)$ is the dissimilarity between observations $\mathbf{x}_m$ and $\mathbf{x}_n$. As the Euclidean distance can not be applied on categorical data, so-called simple matching is used (Kaufman & Rousseeuw, 1990). Let $\mathbf{a} = [a_1, a_2, ..., a_N]$ and $\mathbf{b} = [b_1, b_2, ..., b_N]$ be two observations drawn from the data set. The simple matching dissimilarity measure is given by,

$$\text{dis}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{N} \delta(a_i, b_i), \tag{3}$$

where

$$\delta(a_i, b_i) = \begin{cases} 0, & \text{if } a_i = b_i \\ 1, & \text{if } a_i \neq b_i \end{cases}.$$

The main goal of PAM is to minimise the mean dissimilarity of observations which are closest to observations in the same cluster, compared to the K-means algorithm which minimises the total squared error.

The PAM algorithm consists of two phases (Kumar & Wasan, 2011). The first phase is the "build" phase, in which K (number of clusters) observations are selected for an initial set of medoids, which we call S. This phase follows the following algorithm:

1. Initialise S by randomly choosing K observations as medoids
2. Consider an observation $\mathbf{x}_p \notin S$, which is a candidate to be chosen as new medoid.
3. Consider a different observation $\mathbf{x}_q$, which is not yet in S ($\mathbf{x}_q \notin S$).
4. Calculate the dissimilarity of $\mathbf{x}_q$ with the most similar medoid in the set of medoids S, $\text{dis}(\mathbf{x}_q, .)$.
5. Calculate the dissimilarity of observation $\mathbf{x}_q$ with the potential new medoid $\mathbf{x}_p$, $\text{dis}(\mathbf{x}_q, \mathbf{x}_p)$.
6. Calculate the difference between these two differences, $\text{dis}(\mathbf{x}_q, .) - \text{dis}(\mathbf{x}_q, \mathbf{x}_p)$.
7. If the calculated difference is positive (i.e., $\text{dis}(\mathbf{x}_q, .) > \text{dis}(\mathbf{x}_q, \mathbf{x}_p)$), we know that observation $\mathbf{x}_q$ contributes to the possible selection of observation $\mathbf{x}_p$. Let $D_{qp} = \max\{\text{dis}(\mathbf{x}_q, .) - \text{dis}(\mathbf{x}_q, \mathbf{x}_p), 0\}$.
8. Sum D_{qp} over all possible q: $\sum_{\{q\,:\,\mathbf{x}_q \notin S\}} D_{qp}$.
9. Choose the observation p that maximises $\sum_{\{q\,:\,\mathbf{x}_q \notin S\}} D_{qp}$.
10. Repeat these steps until K medoids have been found, $|S| = K$.

The second phase is called the "swap" phase. In this phase we try to improve S and thus the clustering quality. We consider all pairs $(\mathbf{x}_p, \mathbf{x}_h)$ in which $\mathbf{x}_p$ is chosen as medoid and $\mathbf{x}_h$ is not. We then determine if the clustering quality is improved when swapping $\mathbf{x}_h$ with $\mathbf{x}_p$ in the set of medoids S. The following algorithm is used:

1. Consider observation $\mathbf{x}_q$ that has not yet been selected. We calculate the swap contribution, which we call C_{qph}, as follows:.
   (a) if $\mathbf{x}_q$ is further from $\mathbf{x}_p$ and $\mathbf{x}_h$ compared to the other medoids in S, set $C_{qph} = 0$.
   (b) if $\mathbf{x}_q$ is not further from $\mathbf{x}_p$ compared to all other medoids ($\text{dis}(\mathbf{x}_q, \mathbf{x}_p) = \text{dis}(\mathbf{x}_q, .)$, consider one of two possible situations:

i. $\mathbf{x}_q$ is closer to $\mathbf{x}_h$ than the second closest medoid in S and $\text{dis}(\mathbf{x}_q, \mathbf{x}_h) < \text{dis}_2(\mathbf{x}_q, .)$, where $\text{dis}_2(\mathbf{x}_q, .)$ is the dissimilarity between $\mathbf{x}_q$ and the second most similar medoid in S. If this is the case, then $C_{qph} = \text{dis}(\mathbf{x}_q, \mathbf{x}_h) - \text{dis}(\mathbf{x}_q, \mathbf{x}_i)$.

ii. $\mathbf{x}_q$ is at least as far away from $\mathbf{x}_h$ as the second closest medoid $(\text{dis}(\mathbf{x}_q, \mathbf{x}_h) \geq \text{dis}_2(\mathbf{x}_q, .)$. If this is the case, then $C_{qph} = \text{dis}_2(\mathbf{x}_q, .) - \text{dis}(\mathbf{x}_q, .)$.

(c) if $\mathbf{x}_q$ is further away from $\mathbf{x}_p$ than at least one of the other chosen medoids in S, however, closer to $\mathbf{x}_h$ than to any other medoid in S, then $C_{qph} = \text{dis}(\mathbf{x}_p, \mathbf{x}_h) - \text{dis}(\mathbf{x}_q, .)$.

2. Compute the total result of the swap as $T_{ih} = \sum_j C_{qph}$.

3. Select the pair $(\mathbf{x}_p, \mathbf{x}_h)$ with the lowest T_{ih}.

4. If $T_{ih} < 0$, the swap is performed and the swap phase is again executed from step 1. If $T_{ih} \geq 0$, the objective value cannot be lowered by swapping observations and the algorithm is ended.

One of the drawbacks of applying the PAM algorithm for clustering, is its long computation time, which increases exponentially with the size of the data set. Another drawback is that, similar to K-means, the number of clusters needs to be defined prior to running the algorithm. This can give problems when the optimal amount of clusters is unknown prior to the analysis being applied. To determine which value of k is optimal, the algorithm should be applied on a set of different values of k, which increases the computation time even further.

3.2.3 Principal component analysis

Applying a cluster algorithm on a high-dimensional data set can result in the previously discussed curse of dimensionality (Section 3.2.1). To combat this problem, two methods that reduce the number of dimensions are introduced. The first method is principal component analysis (PCA). PCA reduces the number of dimensions by creating principal components (PC), with each PC being a linear combination of the old variables. These PC's are uncorrelated with each other and explain as much variance as possible.

Before applying PCA, the data matrix $\mathbf{X}$ needs to be standardised. This is done by subtracting the column mean of each observation, resulting in the matrix $\mathbf{W}$. Each element $w_{ij} = x_{ij} - \bar{x}_{.j}$, with w_{ij} and x_{ij} being the i'th observation of variable j from matrix $\mathbf{W}$ and $\mathbf{X}$ respectively and $\bar{x}_{.j} = \sum_{i=1}^{N} x_{ij}$ being the column mean. Using $\mathbf{W}$, we can compute the spectral decomposition of $\mathbf{W}'\mathbf{W}$, which gives us the eigenvalues and eigenvectors of $\mathbf{X}$. The spectral decomposition is shown in Equation (4).

$$\mathbf{W}'\mathbf{W}\mathbf{S} = \mathbf{S}\boldsymbol{\Lambda} \tag{4}$$

Here, $\mathbf{S}$ is a matrix with as its columns the eigenvectors of $\mathbf{X}$. $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of $\mathbf{X}$ on its diagonal, where each eigenvalue corresponds to the eigenvector on the same index. Multiplying the eigenvectors with the original data $\mathbf{X}$ gives the principal components and dividing the corresponding eigenvalue (λ_j) through the total sum of eigenvalues gives the explained variance of that principal component.

The eigenvectors can also be retrieved by use of singular value decomposition. This results in an orthogonal least squares approximation of the centred data. The singular value decomposition is given by:

$$\mathbf{W} = \mathbf{R}\boldsymbol{\Sigma}\mathbf{S}', \tag{5}$$

where $\mathbf{R}$ and $\mathbf{S}$ are orthogonal matrices and $\boldsymbol{\Sigma}$ is the diagonal matrix with the singular values on its diagonal.

Writing $\mathbf{W}'\mathbf{W}$ using Equation (5) and using the orthogonal property of $\mathbf{R}$ and $\mathbf{S}$ ($\mathbf{R}'\mathbf{R} = \mathbf{S}'\mathbf{S} = \mathbf{I}$), we can show the relation between the spectral decomposition and the singular value decomposition. This is shown in Equation (6).

$$\mathbf{W}'\mathbf{W}\mathbf{S} = \mathbf{S}\boldsymbol{\Sigma}\mathbf{R}'\mathbf{R}\boldsymbol{\Sigma}\mathbf{S}'\mathbf{S} = \mathbf{S}\boldsymbol{\Sigma}\boldsymbol{\Sigma} \tag{6}$$

Observing Equations (4) and (6), we see can see that $\boldsymbol{\Sigma}\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$. Thus, the eigenvalues of $\mathbf{X}$ are equal to the squared singular values.

Using the decompositions, we can choose a number of components and define its explained variance by calculating the ratio of the eigenvalues of the chosen components relative to the sum of all eigenvalues. This ratio gives the explained variance and can be used to determine how many components need to be chosen to give a adequate approximation of the data.

However, an alternative approach to choose the number of principal components is by use of Kaiser's criterion (Kaiser, 1960). This criterion states that the number of principal components chosen is decided by choosing the components that have an eigenvalue larger than 1.

3.2.4 Correspondence analysis and multiple correspondence analysis

When coding the data as categorical data, correspondence analysis (CA) is used (Benzécri, 1973).

To explain CA, we use the notation following Van De Velden et al. (2017). We make use of the two-way contingency table $\mathbf{T}$, with non-negative elements. This table cross-tabulates two categorical variables, where the elements give the number of times a combination of the two categories occurred. Create the matrix of fractions, $\mathbf{P}$, is done by dividing each element of $\mathbf{T}$ by the total sum of elements: $\mathbf{P} = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{Q} t_{ij}}\mathbf{T}$, such that the elements of $\mathbf{P}$ sum op to one. CA minimises the following objective function:

$$\phi_{ca}(\mathbf{A},\mathbf{B}) = \left\|\tilde{\mathbf{P}} - \mathbf{D}_r^{\frac{1}{2}}\mathbf{A}\mathbf{B}'\mathbf{D}_c^{\frac{1}{2}}\right\|^2 \tag{7}$$

such that

$$\mathbf{B}'\mathbf{D}_c\mathbf{B} = \mathbf{I}_D.$$

Here $\tilde{\mathbf{P}} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-\frac{1}{2}}$, with $(\mathbf{P} - \mathbf{r}\mathbf{c}')$ being the centred matrix of $\mathbf{P}$ centred symmetrically by row and columns. The row and column sums are given by the vectors $\mathbf{r}$ and $\mathbf{c}$ respectively

($\mathbf{r} = \mathbf{P1}_Q$, $\mathbf{c} = \mathbf{P}'\mathbf{1}_N$, with $\mathbf{1}_L$ being a vector of ones of length L). $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices with the vectors $\mathbf{r}$ and $\mathbf{c}$ on its diagonal. The matrices $\mathbf{A}$ and $\mathbf{B}$ are coordinate matrices of rank D, with D being the dimensionality of the reduced space. A solution of the objective function can be acquired by using singular value decomposition on $\tilde{\mathbf{P}}$, using Equation (5).

If the first D columns of $\mathbf{R}$, $\mathbf{S}$ and $\mathbf{\Sigma}$ are chosen, $\tilde{\mathbf{P}}$ is approximated in D dimensions. This results in the following coordinate matrices:

$$\mathbf{A} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{R\Sigma} \tag{8}$$

and

$$\mathbf{B} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{S}, \tag{9}$$

such that

$$\mathbf{A}'\mathbf{D}_r\mathbf{A} = \mathbf{\Sigma}^2. \tag{10}$$

Combing Equation (5) for $\tilde{\mathbf{P}}$, (8) and (9) gives us the following relation between $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{A} = \mathbf{D}_r^{-\frac{1}{2}}\tilde{\mathbf{P}}\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}\tilde{\mathbf{P}}\mathbf{D}_c^{\frac{1}{2}}\mathbf{B}. \tag{11}$$

Inserting Equation (11) into the objective function stated in Equation (7) results in the objective function:

$$\phi_{ca}'(\mathbf{B}) = \left\| \tilde{\mathbf{P}} - \mathbf{D}_r^{\frac{1}{2}}\mathbf{AB}'\mathbf{D}_c^{\frac{1}{2}} \right\|^2 = \text{tr}\left(\tilde{\mathbf{P}}'\tilde{\mathbf{P}}\right) - \text{tr}\left(\mathbf{B}'\mathbf{D}_c^{\frac{1}{2}}\tilde{\mathbf{P}}'\tilde{\mathbf{P}}\mathbf{D}_c^{\frac{1}{2}}\mathbf{B}\right) \tag{12}$$

such that

$$\mathbf{B}'\mathbf{D}_c\mathbf{B} = \mathbf{I}_D,$$

with Tr($*$) indicating the trace operator, which gives the sum of the diagonal elements of the matrix on which it is operated. We can see that minimising $\phi_{ca}'(\mathbf{B})$ is equivalent to maximising $-\phi_{ca}'(\mathbf{B})$. Removing the constant term tr $\left(\tilde{\mathbf{P}}'\tilde{\mathbf{P}}\right)$ in Equation (12) result in the objective function stated in Equation (13), which needs to be maximised.

$$\phi_{ca}'(\mathbf{B}) = \text{tr}\left(\mathbf{B}'\mathbf{D}_c^{\frac{1}{2}}\tilde{\mathbf{P}}'\tilde{\mathbf{P}}\mathbf{D}_c^{\frac{1}{2}}\mathbf{B}\right) \tag{13}$$

such that

$$\mathbf{B}'\mathbf{D}_c\mathbf{B} = \mathbf{I}_D.$$

This objective function needs to be maximised over $\mathbf{B}$, which results in the between row variance being maximised.

CA can also be applied more than two categorical variables. One method that is applicable on three or more categorical variables is multiple correspondence analysis (MCA, Greenacre (1984)). This method uses the indicator matrix $\mathbf{Z}$ (Section 3.1.3). Applying CA on the indicator matrix results in the MCA analysis.

3.2.5 Correspondence analysis of ratings

Greenacre (1984) showed that to take the bipolar and absolute nature of the ratings into account, correspondence analysis can be applied on the doubled data matrix (Section 3.1.2). This doubling of the data creates a symmetry between the two poles, which makes the CA invariant to the scale direction. Applying CA to the doubled matrix is called correspondence analysis of ratings (CAr).

Take a pair of doubled columns, indexed by j_+ and j_-, which contain the values y_{ij} and $t_j - y_{ij}$ respectively, with t_j being the upper bound of rating variable j with the rating variable j ranging from 0 to t_j, $j = 1, ..., Q$ and $i = 1, ..., N$. This data matrix has row sums equal to the constant $t_{.} = \sum_{j=1}^{Q} t_j$, which makes it such that the row masses of the matrix are all equal to $\frac{1}{N}$. The column sums of columns j_+ and j_-, however, are not equal, but are $y_{.j}$ and $Nt_j - y_{.j}$ respectively. Here $y_{.j} = \sum_{i=1}^{N} y_{ij}$. This makes the column masses c_{j_+} and c_{j_-} equal to $\frac{\bar{y}_j}{t_{.}}$ and $\frac{(t_j - \bar{y}_j)}{t_{.}}$ respectively, with $\bar{y}_j = \frac{y_{.j}}{N}$.. If t_j is equal for all j, the sum of column masses equals $c_{j_+} + c_{j_-} = \frac{1}{Q}$. Due to the fact that the two columns of a doubled pair sum to a constant, it is clear to see that the centroids of this doubled pair of points lie at the origin of the CA graph. The points j_+ and j_- are balanced at the origin, with the "lighter" point laying relatively further away from the origin. This means that the point that is closer to the origin has a higher association with the rating variable. Thus, if the CA point of the positive variable j_+ lies closer to the origin, variable j has a relatively high rating.

Furthermore, we can show that the distances from the CA points corresponding with j_+ and j_- and the origin, d_{j_+} and j_{q_-} respectively, are equal to the variation of the respective columns in the doubled matrix. This implies that

$$d_{j_+} = \frac{s_j}{\bar{y}_j}, \tag{14}$$

which is the variation (standard deviation divided by the mean) and

$$d_{j_-} = \frac{s_j}{(t_j - \bar{y}_j)}, \tag{15}$$

with $s_j = \sqrt{(\sum_{i=1}^{N} y_{ij}^2 - N\bar{y}_j^2)}$.

Greenacre also shows that the correspondence analysis of the matrix of doubled bipolar data is equivalent to applying PCA on the matrix $\mathbf{Y}^*$:

$$\mathbf{Y}^* = \mathbf{Y} \cdot \text{diag}(\sqrt{\boldsymbol{\xi}}), \tag{16}$$

where diag($\sqrt{\boldsymbol{\xi}}$) is a $Q \times Q$ matrix with vector $\sqrt{\boldsymbol{\xi}}$ on the diagonal axis and zeros everywhere else. $\boldsymbol{\xi}$ is a vector of length Q with

$$\xi_j = \frac{(t_j/t_{.})}{\bar{y}_j(t_j - \bar{y}_j)}. \tag{17}$$

In Equation (17) $\bar{y}_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij}$.

Applying PCA on $\mathbf{Y}^*$ sets the rows of $\mathbf{Y}^*$ in an Euclidean space with masses $1/N$ and the squared interpoint distances between rows i and i^* of $\sum_j (y_{ij}^* - y_{i^*j}^*)^2$. These are the same masses and relative positions as the rows of the doubled matrix (Greenacre, 1984). Applying this scaling to the data, makes it so that PCA can be applied and give the same results as correspondence analysis on the doubled matrix, however, only for the undoubled data. To retrieve the doubled data points from the PCA analysis, we can use Equation 15 in combination with the fact that the doubled points are balanced around the origin to calculate the doubled points.

3.2.6 Tandem analysis

Where the K-means and PAM analyses are applied on the complete data set, the third clustering strategy that we consider, the so-called tandem analysis, is a combination of dimension reduction as well as clustering. We use two different variations of the tandem analysis, because of the different types of data that are being used.

For the raw rating data, we use the tandem analysis of PCA followed by K-means, as discussed in the previous sections (Vichi & Kiers, 2001). This method is widely used by researchers because it is easy to apply and it is fairly straightforward. First, the number of dimensions is reduced by applying PCA on the full data set and choosing the number of components following Kaiser's criterion or evaluating the explained variance. After choosing a sufficient number of components, by using the total explained variance or Kaiser's criterion (Section 3.2.3), the K-means algorithm is applied on the chosen principal components.

For doubled rating data, we apply tandem analysis of doubled CA and K-means, which is equivalent to applying PCA on the matrix $\mathbf{Y}^*$ (Section 3.2.4) followed by K-means.

Finally, the tandem analysis of MCA and K-means is applied on categorical data, using the indicator matrix $\mathbf{Z}$.

3.2.7 Reduced K-means

The tandem analysis, as discussed in the previous section, results in a clustering as well as a dimension reduction. However, these two methods are done separately and both optimise a different objective function. This could lead to non-optimal results. When PCA retrieves components that do explain a lot of variance, but are not relevant to the clustering structure, the tandem analysis could fail to retrieve the underlying clustering structure. Therefore, using a method that optimises one objective function, that both clusters the data as well as reduces the number of dimensions, could be of large interest. Reduced K-means (RKM) is such a method (De Soete & Carroll, 1994).

RKM combines both the cluster membership matrix and the reduced dimensions, making it so that both of these matrices are optimised simultaneously (Timmerman et al., 2010). The loss function that is being minimised is,

$$F(\mathbf{U}, \mathbf{F}, \mathbf{A}) = \|\mathbf{X} - \mathbf{UFA}'\|^2, \tag{18}$$

where $\mathbf{X}$ is the $N \times Q$ data matrix, $\mathbf{U}$ is the cluster membership matrix with dimensions $N \times K$, $\mathbf{F}$ is the $K \times D$ matrix containing the centroids of the clusters (with D the number of reduced dimensions) and $\mathbf{A}$ is the $Q \times D$ loading matrix. Here $\|.\|$ is the Euclidean norm. Which is equivalent to the Euclidean distance stated in Equation (2) on page 12, however, the double sum is taken over all elements of the matrix.

Following the notation of Yamamoto & Hwang (2014), we rewrite the loss function by using the solution for the cluster means: $\mathbf{F} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}$. Inserting this in the loss function of Equation (18) results in

$$F(\mathbf{U}, \mathbf{A}) = \|\mathbf{X} - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}\mathbf{A}'\|^2 = \|\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}'\|^2, \tag{19}$$

where $\mathbf{P} = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$. That is, $\mathbf{P}$ is the influence matrix on the space which is spanned by the columns of the cluster membership matrix $\mathbf{U}$. The loss function from Equation (19) can be decomposed, using that the loading matrix $\mathbf{A}$ can be rotated. Using that

$$\begin{aligned} \operatorname{tr}\left[(\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}')'(\mathbf{X}\mathbf{A}\mathbf{A}' - \mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}')\right] &= \operatorname{tr}\left[(\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}')'(\mathbf{X}\mathbf{A} - \mathbf{P}\mathbf{X}\mathbf{A})\mathbf{A}'\right] \\ &= \operatorname{tr}\left[(\mathbf{A}'(\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}')'(\mathbf{X}\mathbf{A} - \mathbf{P}\mathbf{X}\mathbf{A})\right] \\ &= \operatorname{tr}\left[((\mathbf{X}\mathbf{A} - \mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{A})'(\mathbf{X}\mathbf{A} - \mathbf{P}\mathbf{X}\mathbf{A})\right] \\ &= \operatorname{tr}\left[((\mathbf{X}\mathbf{A} - \mathbf{X}\mathbf{A})'(\mathbf{X}\mathbf{A} - \mathbf{P}\mathbf{X}\mathbf{A})\right] \\ &= 0 \end{aligned} \tag{20}$$

we can decompose the objective function as follows:

$$\begin{aligned} F(\mathbf{U}, \mathbf{A}) = \|\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}'\|^2 &= \|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}' + \mathbf{X}\mathbf{A}\mathbf{A}' - \mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}'\|^2 \\ &= \|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}'\|^2 + \|(\mathbf{X}\mathbf{A} - \mathbf{P}\mathbf{X}\mathbf{A})\mathbf{A}'\|^2 \\ &\quad + 2 \cdot \operatorname{tr}\left[(\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}')'(\mathbf{X}\mathbf{A}\mathbf{A}' - \mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}')\right] \\ &= \|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}'\|^2 + \|\mathbf{X}\mathbf{A} - \mathbf{P}\mathbf{X}\mathbf{A}\|^2. \end{aligned} \tag{21}$$

We can see that the first part of Equation (21) corresponds to the PCA objective function. The second part of the Equation equals the objective function of K-means in the reduced space. Minimising Equation (21) is done by means of an alternating least-squares (ALS) algorithm is used (De Leeuw et al., 1976). This algorithm alternates between two steps until convergence. In the first step of the algorithm the loss function is minimised over $\mathbf{U}$, while keeping $\mathbf{A}$ fixed. Looking at Equation 21, we can see that minimising $F(\mathbf{U}, \mathbf{A})$ over $\mathbf{U}$ is done by minimising the second term of this equation, because the first term does not contain $\mathbf{U}$. Minimising this term can be done by applying standard K-means (Section 3.2.1).

The second step of the ALS algorithm minimises the loss function over $\mathbf{A}$, keeping $\mathbf{U}$ fixed. Using Equation (19) in combination with the trace operator, the loss function can be rewritten as:

$$F(\mathbf{U}, \mathbf{A}) = \|\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}'\|^2 = \operatorname{tr}(\mathbf{X}'\mathbf{X}) - \operatorname{tr}(\mathbf{A}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{A}). \tag{22}$$

We can quickly see that minimising the loss function $F(\mathbf{U},\mathbf{A})$ is equivalent to maximising $-F(\mathbf{U},\mathbf{A})$, resulting in maximisation of the between cluster variance in the reduced space: $\mathrm{tr}(\mathbf{A}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{A})$ over $\mathbf{A}$. This is done by solving the eigenvalue problem using singular value decomposition:

$$\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{A} = \mathbf{A}\mathbf{\Sigma}, \tag{23}$$

where $\mathbf{\Sigma}$ is an $D \times D$ diagonal matrix.

3.2.8 Factorial K-means

The second simultaneous method that we use is closely related to RKM. The so-called factorial K-means (FKM) analysis is also a method that optimises one objective function for both clustering the data as well as reducing the number of dimensions (Vichi & Kiers, 2001).

The main difference between FKM and RKM is the loss function (Timmerman et al., 2010). The FKM loss function is stated as

$$F(\mathbf{U},\mathbf{F},\mathbf{A}) = \|\mathbf{X}\mathbf{A}\mathbf{A}' - \mathbf{U}\mathbf{F}\mathbf{A}'\|^2 = \|\mathbf{X}\mathbf{A} - \mathbf{U}\mathbf{F}\|^2. \tag{24}$$

Minimising this loss function makes it such that $\mathbf{U}\mathbf{F}\mathbf{A}'$ needs to be as close to $\mathbf{X}\mathbf{A}\mathbf{A}'$ as possible.

Using the notation for $\mathbf{P}$ as used in the previous section ($\mathbf{P} = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$) and using that $\mathbf{F} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}$ we can rewrite Equation (24) as:

$$F(\mathbf{U},\mathbf{A}) = \|\mathbf{X}\mathbf{A} - \mathbf{P}\mathbf{X}\mathbf{A}\|^2 = \mathrm{tr}(\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A}) - \mathrm{tr}(\mathbf{A}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{A}). \tag{25}$$

Again, we can use an ALS algorithm to minimise Equation 25. In the first step, we minimise the loss function over $\mathbf{U}$, for fixed $\mathbf{A}$. This can be done by using K-means, equivalent to RKM. In the second step of the ALS algorithm, we minimise $F(\mathbf{U},\mathbf{A})$ over $\mathbf{A}$, while keeping $\mathbf{U}$ fixed. This is equivalent to solving the following eigenvalue problem:

$$\mathbf{X}'(\mathbf{P} - \mathbf{I}_N)\mathbf{X}\mathbf{A} = \mathbf{A}\mathbf{\Sigma}, \tag{26}$$

where $\mathbf{\Sigma}$, again, is a $D \times D$ diagonal matrix and $\mathbf{I}_N$ denotes an $N \times N$ indicator matrix.

As we can see, RKM minimises the sum of squared distances between the centroids in the reduced space ($\mathbf{UFA}$') and the data ($\mathbf{X}$), whereas FKM minimises the sum of squared distances between the centroids in the reduced space and the data projected onto the reduced subspace ($\mathbf{X}\mathbf{A}\mathbf{A}'$). Thus, the main difference between RKM and FKM is that RKM reconstructs the whole data set with only the centroids being in the reduced space, meaning that the sum of squared distances between the observations and the centroids located in the reduced space is minimised. However, FKM minimises the sum of squared distances with the centroids as well as the observations projected in the reduced space. That is, FKM optimised the within-cluster sum of squared distances in the reduced space (Timmerman et al., 2010; Yamamoto & Hwang, 2014).

3.2.9 Cluster correspondence analysis

For the same reason as with PCA, we cannot apply RKM and FKM on a categorical data set. Therefore we use a different method when analysing categorical data, namely, cluster correspondence analysis (Van De Velden et al., 2017).

For this method, we first need to construct the super indicator matrix, $\mathbf{Z}$, as in Section 3.1.3. The matrix has a dimensionality of $N \times T$, where $T = \sum_{j=1}^{Q} q_j$ and Q the number of variables, thus all possible answers on all questions. In our case, we can see the ratings as categories. The cluster memberships can also be seen as a categorical variable, which leads to the cluster membership indicator matrix $\mathbf{U}$, with dimensionality $N \times K$. Using these two matrices, the matrix cross-tabulating cluster membership with the rating variables is constructed, $\mathbf{F} = \mathbf{U}'\mathbf{Z}$.

Applying correspondence analysis on $\mathbf{F}$, the optimal scaling values for the clusters and categories, which are the rows and columns of $\mathbf{F}$ respectively, are retrieved, with the variance from observations of the same cluster being minimised and the variance of observations of different clusters being maximised.

CCA is explained following the notation of Van De Velden et al. (2017). First we define the matrix of fractions by dividing each element of $\mathbf{F}$ by the total sum of elements, i.e. $\mathbf{P} = \frac{1}{NQ}\mathbf{F}$. Evaluating $\mathbf{P} - \mathbf{r}\mathbf{c}'$ leads to:

$$\mathbf{P} - \mathbf{P}\mathbf{1}_T\mathbf{1}_K'\mathbf{P} = \frac{1}{NQ}\left(\mathbf{F} - \frac{1}{NQ}\mathbf{F}\mathbf{1}_T\mathbf{1}_K'\mathbf{F}\right) = \frac{1}{NQ}\left(\mathbf{U}'\mathbf{Z} - \frac{1}{N}\mathbf{U}'\mathbf{1}_N\mathbf{1}_N'\mathbf{Z}\right) = \frac{1}{NQ}\mathbf{U}'\mathbf{M}\mathbf{Z}. \quad (27)$$

Here $\mathbf{r}$ and $\mathbf{c}$ have the same definition as in Section 3.2.4 (with $\mathbf{P}$ having a different dimensionallity) and $\mathbf{M} = \mathbf{I}_N - \mathbf{1}_N\mathbf{1}_N'/N$.

We now define $\mathbf{D}_K = \mathbf{U}'\mathbf{U}$ as a diagonal matrix of cluster sizes, as well as the diagonal matrix $\mathbf{D}_\mathbf{z}$ such that $\mathbf{D}_z\mathbf{1}_T = \mathbf{Z}'\mathbf{1}_n$. Using this notation, the CA objective function as stated in Equation (13) (page 17) becomes

$$\phi_{CCA}(\mathbf{U}, \mathbf{B}) = \frac{1}{NQ^2}\text{tr}\left(\mathbf{B}'\mathbf{Z}'\mathbf{M}\mathbf{U}\mathbf{D}_K^{-1}\mathbf{U}'\mathbf{M}\mathbf{Z}\mathbf{B}\right), \quad (28)$$

subject to

$$\frac{1}{NQ}\mathbf{B}'\mathbf{D}_z\mathbf{B} = \mathbf{I}_D.$$

Here D is the number of dimensions of the CA solution. The value of D must be chosen subject to the constraint $D \leq \min(K-1, T-1)$.

Defining $\mathbf{B}^* = \frac{1}{\sqrt{NQ}}\mathbf{D}_z^{\frac{1}{2}}\mathbf{B}$, we can rewrite the objective function in Equation (28) as

$$\phi_{CCA}(\mathbf{U}, \mathbf{B}^*) = \frac{1}{Q}\text{tr}\left(\mathbf{B}^{*\prime}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{Z}'\mathbf{M}\mathbf{U}\mathbf{D}_K^{-1}\mathbf{U}'\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^*\right) \quad (29)$$

subject to

$$\mathbf{B}^{*\prime}\mathbf{B}^* = \mathbf{I}_d.$$

For a fixed cluster membership matrix $\mathbf{U}$, we can get the solution of $\mathbf{B}^*$ from the eigenvalue decomposition

$$\frac{1}{Q}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{Z}'\mathbf{M}\mathbf{U}\mathbf{D}_K^{-1}\mathbf{U}'\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}} = \mathbf{B}^*\boldsymbol{\Sigma}^2\mathbf{B}^{*'}. \tag{30}$$

Using the solution $\mathbf{B}^*$ we can calculate $\mathbf{B} = \sqrt{NQ}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^*$. Then, to optimise Equation (29), we need to determine the optimal cluster allocation $\mathbf{U}$. For a fixed $\mathbf{B}^*$ the optimisation can be rewritten as a K-means clustering method. This means that maximising Equation (29) with respect to $\mathbf{U}$ is equivalent to minimising

$$\phi'_{CCA}(\mathbf{U},\mathbf{F}) = \left\|\sqrt{\frac{N}{Q}}\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^* - \mathbf{U}\mathbf{F}\right\|^2, \tag{31}$$

where $\mathbf{F}$ is a matrix containing the cluster means.

We now rewrite the subject coordinates as

$$\mathbf{Y} = \sqrt{\frac{N}{Q}}\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^* \tag{32}$$

and thus Equation (31) becomes:

$$\phi'_{CCA}(\mathbf{U},\mathbf{F}) = \|\mathbf{Y} - \mathbf{U}\mathbf{F}\|^2. \tag{33}$$

Minimising Equation (33) over $\mathbf{F}$ leads to the following solution:

$$\mathbf{F}^* = \left(\mathbf{U}'\mathbf{U}\right)^{-1}\mathbf{U}'\mathbf{Y} = \mathbf{D}_K^{-1}\mathbf{U}'\mathbf{Y}. \tag{34}$$

Inserting the solution $\mathbf{F}^*$ into the K-means function from Equation (33) leads to:

$$\begin{aligned}\phi'_{CCA}(\mathbf{U},\mathbf{F}^*) &= \|\mathbf{Y} - \mathbf{U}\mathbf{F}^*\|^2 \\ &= \mathrm{tr}(\mathbf{Y}'\mathbf{Y}) + \mathrm{tr}(\mathbf{F}'\mathbf{D}_K\mathbf{F}) - 2\cdot\mathrm{tr}(\mathbf{F}'\mathbf{U}'\mathbf{Y}) \\ &= \mathrm{tr}(\mathbf{Y}'\mathbf{Y}) + \mathrm{tr}(\mathbf{Y}'\mathbf{U}\mathbf{D}_K^{-1}\mathbf{D}_K\mathbf{D}_K^{-1}\mathbf{U}'\mathbf{Y}) - 2\cdot\mathrm{tr}(\mathbf{Y}'\mathbf{U}\mathbf{D}_K^{-1}\mathbf{U}'\mathbf{Y}) \\ &= \mathrm{tr}(\mathbf{Y}'\mathbf{Y}) - \mathrm{tr}(\mathbf{Y}'\mathbf{U}\mathbf{D}_K^{-1}\mathbf{U}'\mathbf{Y}).\end{aligned} \tag{35}$$

This shows that minimising the K-means objective over $\mathbf{U}$ and $\mathbf{F}$ is equivalent to maximising the term on the right-hand side of Equation (35). This can be rewritten as:

$$\mathrm{tr}(\mathbf{Y}'\mathbf{U}\mathbf{D}_K^{-1}\mathbf{U}'\mathbf{Y}) = N\cdot\mathrm{tr}\left(\frac{1}{Q}\mathbf{B}^{*'}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{Z}'\mathbf{M}\mathbf{U}\mathbf{D}_K^{-1}\mathbf{U}'\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^*\right). \tag{36}$$

This results in the function only needing to be optimised over $\mathbf{U}$. Thus, for a fixed $\mathbf{B}^*$, we can find an optimal $\mathbf{U}$ by applying the K-means algorithm on $\mathbf{Y}$. With the new cluster allocation ($\mathbf{U}$), the optimal scaling values ($\mathbf{B}^*$) are updated and then $\mathbf{U}$ is calculated again. Repeating this process leads to the following algorithm:

1. Generate the cluster membership matrix $\mathbf{U}$. This is done by assigning observations to clusters randomly.
2. Find the optimal scaling values $\mathbf{B}^*$ using Equation (30).
3. Calculate $\mathbf{Y}$ using Equation (32) and apply K-means on these coordinates to receive a more optimal cluster membership matrix $\mathbf{U}$.
4. Repeat steps 2 and 3 until convergence of $\mathbf{U}$ (and hence $\mathbf{Y}$ and $\mathbf{F}$).

This algorithm guarantees convergence, because the value of the objective function decreases (or stays constant) with each step. To combat the problem of the solution being a local minimum, more random starts can be used.

3.3 Determining the number of clusters

When applying the previously discussed cluster algorithms, the number of clusters needs to be prespecified. Determine the number of clusters can be done using many different methods. One of these methods is the elbow method (Thorndike, 1953). The elbow method is based on comparing the within-cluster sum of squares for different values of K. Comparing these sum of squares is done by plotting a line graph where each within-cluster sum of squares is plotted for a predefined set of number of clusters. K is selected by choosing the number of clusters for which the decrease in within-cluster sum of squares is levelling off, which appears as an "elbow" in the graph.

The intuition behind the elbow method is that as K increases, the within-cluster sum of squares decreases, as the data set is divided in smaller and more similar clusters. However, when K keeps increasing, at a certain point the decrease in within-cluster sum of squares becomes smaller. Increasing K in this situation leads to overfitting. Therefore, the value of K depicted by the elbow in the graph is chosen, as this value provides a reasonable trade-off between bias and variance in the clustering.

3.4 Combinations of data coding and cluster analyses

In the previous section, we discussed different types of data coding as well as cluster analyses. However, not all cluster analyses are applicable on each type of data coding. In this section we discuss which cluster analyses are applied on which type of data. For each type of data coding we use a cluster analyses on the complete data set, a method that successively applies dimension reduction and clustering methods and a method that simultaneously reduces the dimensionality of the data as well as allocates the data in clusters.

The first type of data that is used is the raw rating data. The first method applied on this raw data is regular K-means. Next we use a tandem analysis, which successively applies PCA and K-means. The simultaneous methods used are reduced K-means and factorial K-means.

Secondly, we use the doubled data. We first apply the K-means algorithm on the doubled data set. Applying K-means on the doubled data set is equivalent to applying it on the raw data set. This is due to the fact that the data on the reversed scale has the exact same distance between individuals compared to the raw data points. This means that when the mean on the reversed scale is calculated, the mean vector is similar to the mean vector of the raw data, however, on the reversed scale. When calculating the optimal objective value over the raw rating data (first half of columns), the value over the doubled data set would also lead to the lowest objective value, because the last half of the centres are equally far away from the reversed points. This means that applying K-means on the doubled data should lead to the exact same clustering, only with double the value of the objective function. This, however, does not have to be the case when both objective functions are optimised to a different (local) optimum. The tandem method used on the doubled data is CA of ratings, as discussed in Section 3.2.5. This is CA applied on the doubled rating scale data matrix. For simultaneously clustering and reducing the number of dimensions of the doubled rating data, we again use RKM and FKM, however, the raw data is used and standardised as defined in Equation (17) (Section 3.2.5).

Lastly, the data is coded as categorical data using the super indicator matrix **Z**. The first clustering method applied on this data is the partitioning around medoids algorithm. Secondly, we use the tandem analysis of MCA and K-means. Lastly, we apply cluster correspondence analyses on the categorical data.

These combinations of data coding and clustering result in eleven different cluster analyses. A summary of the methods is given in Table 1, which shows how the different methods are implemented.

Table 1: Summary of the eleven clustering methods

Data Coding	Method	Summary
Raw	K-means	K-means, as described in Section 3.2.1, applied on the raw data set.
	Tandem	A tandem analysis, where first PCA (Section 3.2.3) is applied on the raw data to construct principal components. Thereafter, K-means (Section 3.2.1) is applied to cluster the data in the reduced space.
	RKM	Reduced K-means applied on the raw data set.
	FKM	Factorial K-means applied on the raw data set.
Doubled	K-means	K-means, as described in Section 3.2.1, applied on the doubled data set.
	Tandem	A tandem analysis, where first PCA is applied on the standardised data set, following Equation (17) in Section 3.2.5, making it equivalent to CA of ratings (Greenacre, 1984). Thereafter, K-means is applied on the principal components in the reduced space.
	RKM	Reduced K-means applied on the standardised data set, following Equation (17) in Section 3.2.5.
	FKM	Factorial K-means applied on the standardised data set, following Equation (17) in Section 3.2.5.
Categorical	PAM	Partitioning around medoids algorithm (Section 3.2.2) applied on the super indicator matrix **Z** (Section 3.1.3).
	Tandem	A tandem analysis, where first MCA is applied on the super indicator matrix **Z** (Section 3.2.9). Thereafter, K-means is applied on the principal components in the reduced space.
	CCA	Cluster Correspondence Analysis applied using the super indicator matrix **Z**, following the procedure stated in Section 3.2.9.

4 Simulation Study

In this research, both simulated and real data sets are utilised. This section focuses on the analysis of the simulated data sets. We evaluate the performance of eleven different clustering methods, which can be grouped in three groups: raw clustering, successive dimension reduction and clustering and simultaneous dimension reduction and clustering. By applying these methods to simulated data, we are able to identify their relative strengths and weaknesses. Modifying the data generating process allows us to further investigate the specific scenarios in which each method performs optimally.

4.1 Data generating process

The simulated data sets are drawn from a Gaussian mixture model with pre-specified overlap characteristics, following the notation of Melnykov et al. (2012). The Gaussian mixture model consists of the weighted sum of several Gaussian densities. The Equation is given by

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^{K} \pi_k \cdot g(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{37}$$

Here K is the number of Gaussian densities, $\mathbf{x}$ is a data vector of length Q, π_k is the mixture proportion of mixture $k = 1, ..., K$ and $g(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian density for mixture k. The densities are each of a Q-variate Gaussian nature, with the function given by the following Equation,

$$g(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{Q/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right\}. \tag{38}$$

Here $\boldsymbol{\mu}_k$ is a vector containing the means of mixture k, $\boldsymbol{\Sigma}_k$ is the covariance matrix of mixture k and the mixture proportions satisfy the condition $\sum_{k=1}^{K} \pi_k = 1$.

To draw from the Gaussian distribution, we need the parameters ($\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$) to be specified. Therefore, we let $\boldsymbol{\mu}_k$ be drawn K times from a uniform Q-variate hypercube with bounds set to $[0, 1]$. $\boldsymbol{\Sigma}_k$ is drawn from a Wishart distribution. The Wishart distribution has a parameter Q and $Q + 1$ degrees of freedom.

When simulating noise variables, variables are drawn from a uniform hypercube with bounds $[0, 1]$. This ensures the independent nature of the noise variables relative to the underlying clustering structure.

The overlap between the K different Gaussian densities is controlled as well. This is done by means of the pairwise overlap. Let $\mathbf{X}$ be distributed as a finite mixture model $p(\mathbf{x}|\lambda)$. The overlap between the density of cluster k_1 and k_2 is defined as $\omega_{k_1k_2} = \omega_{k_1|k_2} + \omega_{k_2|k_1}$, where $\omega_{k_1|k_2}$ is the probability of the misclassification that $\mathbf{X}$ originates from k_1 but was assigned to k_2. $\omega_{k_2|k_1}$ is defined equivalently. This leads to the following equation:

$$\omega_{k_2|k_1} = \Pr\left[\pi_{k_1} g(\mathbf{x}|\boldsymbol{\mu}_{k_1}, \boldsymbol{\Sigma}_{k_1}) < \pi_{k_2} g(\mathbf{x}|\boldsymbol{\mu}_{k_2}, \boldsymbol{\Sigma}_{k_2}) \mid \mathbf{X} \sim N_p(\boldsymbol{\mu}_{k_1}, \boldsymbol{\Sigma}_{k_2})\right] \tag{39}$$

If $\mathbf{\Sigma}_k$ is multiplied by a constant c (with $c > 0$), it leads to the clusters being inflated if $c > 1$ or deflated if $c < 1$. We can use this constant to reach a pre-defined level of overlap $\omega_{k_1 k_2}(c)$ (Maitra & Melnykov, 2010). This function does not have to be monotone increasing, however, it does benefit from monotonicity in most cases. In the case that the monotonicity is not respected, a new mixture is simulated.

Maitra & Melnykov (2010) show an algorithm used to simulate Gaussian mixture models using a pre-defined value of average overlap: $\bar{\omega}$. The algorithm works as follows:

1. Generate K mean vectors, covariance matrices and mixture proportions. Compute the limiting average overlap $\bar{\omega}^{\infty}$. If $\bar{\omega} < \bar{\omega}^{\infty}$ proceed to step 2, otherwise, start step 1 over.

2. Calculate all pairwise overlaps and the estimate of $\hat{\bar{\omega}}$. If $\hat{\bar{\omega}}$ and $\bar{\omega}$ do not differ significantly, the chosen parameters are sufficient.

3. The scalar c is chosen using root-finding methods, such that $\hat{\bar{\omega}}(c)$ and $\bar{\omega}(c)$ do not differ significantly.

To transform the continuous data, drawn from the Gaussian mixture model, to rating data, cutoff points are chosen. The cutoff points are evenly spread between the upper and lower bound of the mean vector. If the mean vector is randomly drawn from the interval $[l, u]$, with l and u discrete, and the rating data has a point scale of q_j for variable j, the cutoff points are: l, $l + \frac{(u-l)}{q_j}$, $l + \frac{2 \cdot (u-l)}{q_j}$, ... , $l + \frac{(q_j - 1) \cdot (u-l)}{q_j}$, u. This creates q_j intervals, where the values from l to u are assigned to the intervals in increased order. When drawing from the Gaussian mixture model, it can occur that the random mean is drawn close to one of the boundaries, resulting in values being drawn outside the $[l, u]$ interval. When this is the case, to the values drawn lower than l, we assign l as their rating value. Values above u get u as their rating value.

4.2 Experimental design

In this simulation study, we fix the number of observations at $I = 1000$ and utilise a seven-point rating scale for all variables ($q_j = 7,\ j = 1, ..., Q$).

Three different number of questions, $Q \in [3, 6, 20]$, and two different sizes of noise variables, $N \in [0, 3Q]$, are used. This means that when noise is present, we have $Q + N \in [12, 24, 80]$. The reason for using noise variables is to more accurately mimic real world data, in which it is not uncommon to have variables that are not significantly correlated to the underlying cluster allocation (Dave, 1991).

The primary motivation for manipulating the number of questions and noise variables in this study is to investigate the impact of high-dimensional data on clustering performance. Specifically, we are interested in evaluating the ability of clustering algorithms to accurately identify the underlying clustering structure of the data when the number of variables is large. To assess the ability of different clustering methods to capture the original clustering structure, we use clustering on the whole data set as well as applying a tandem and a simultaneous clustering approach. The tandem approach involves optimising two objective functions successively: one to reduce the number of dimensions and one to cluster the data in the space reduced using the first method. In contrast, the simultaneous approach reduces the number of dimensions and clusters the data using one objective function.

We expect the tandem approach to be more sensitive to the presence of noise variables, compared to a simultaneous approach, as these variables may not contribute significantly to the clustering structure, but could still be included in the reduced space. In contrast, the simultaneous approach should perform better in the presence of noise variables, as the objective function evaluates the optimal dimension reduction as well as the optimal clustering. Overall, we anticipate that the performance of both clustering methods is affected by the number of questions and noise variables in the data, with a larger number of variables potentially leading to overfitting and worse clustering performance (Mathivanan et al., 2019).

The number of clusters are set to $K \in [4, 8]$. Four clusters are chosen following Van De Velden et al. (2017). Additionally, we decided to add eight clusters as well. This is mainly done because of the large number of variables in the situation of six and twenty active variables. As the number of reduced dimensions in the RKM and FKM analyses has a maximum of $D = \min(K - 1, Q - 1)$, when $K = 4$, it would not be possible to include all active variables in the reduced space in the situation of $Q \in [6, 20]$. Adding $K = 8$ makes this possible for the situation of six active variables, however, when taking twenty active variables into account, it is still not possible to include all active variables in the reduced space. For this to be possible, we should include $K = 21$ as well, however, this probably leads to overfitting of the model.

For the mixing proportions π_k, there are two different situations we have to take into consideration, one where the proportions are of equal size (balanced) and one where the proportions are not (unbalanced). In the case of balanced data, each mixing proportion $\pi_k = \frac{1}{K}$ for $k = 1, ..., K$. When the data is unbalanced, we choose the mixing proportions as $\Pi_{K=4}$ = [0.55, 0.25, 0.14, 0.06], following Van De Velden et al. (2017). We choose $\Pi_{K=8}$ = [0.31, 0.23, 0.15, 0.11, 0.08, 0.06, 0.04, 0.02], which is created by interpolating between each adjacent pair of values in $\Pi_{K=4}$. The last value (0.06) is interpolated with 0. This leads to the vector $\Pi'_{K=8}$ = (0.55, 0.40, 0.25, 0.19, 0.14, 0.10, 0.06, 0.03). We then divide the whole vector by its total sum, $\Pi_{K=8} = \frac{1}{1.73} \cdot \Pi'_{K=8}$,such that the proportions add up to 1. These proportions are chosen such that the clusters are significantly different in size. The reasoning behind picking these differently balanced proportions is because research has shown that most of the clustering methods are not optimal when performed on unbalanced data (Fränti & Sieranoja, 2018). Our aim is to inspect whether the effect of using unbalanced data is larger or smaller between the different clustering methods.

Two different values of average overlaps are chosen. Either the overlap is small, $\bar{\omega} = 0.001$,or the overlap is large, $\bar{\omega} = 0.100$. Making the average overlap small gives very spread out distributions. Using the clustering algorithms on these spread out distributions should result in more accurate clustering. Increasing the average overlap makes the distributions overlap more and the cluster distribution should thus be more difficult to retrieve (Melnykov, 2016). We would like to inspect if this effect is as large between the different methods.

4.3 Performance measure

Evaluating how well the different clustering methods perform can be done through performance measures. The performance measure used in this research is the adjusted rand index (ARI) (Hubert & Arabie, 1985). The ARI gives an index that shows to what extend two different clustering results (P^* and P) of the same data set are similar. The following Equation is used:

$$\mathrm{ARI}(P^*, P) = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]}, \tag{40}$$

where

$$a = \frac{\sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2 - N}{2},$$

$$b = \frac{\sum_{r=1}^{R} t_{r.}^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2}{2},$$

$$c = \frac{\sum_{c=1}^{C} t_{.c}^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2}{2},$$

$$d = \frac{\sum_{r=1}^{R} \sum_{c=1}^{C} t_{rc}^2 + N^2 - \sum_{r=1}^{R} t_{r.}^2 - \sum_{c=1}^{C} t_{.c}^2}{2},$$

where R is the number of clusters in the first clustering and C in the second clustering, t_{rc} is the number of observations clustered in the r'th subset of clustering R and the c'th subset of clustering C, $t_{r.} = \sum_{c=1}^{C} t_{rc}$ and $t_{.c} = \sum_{r=1}^{R} t_{rc}$ (Steinley, 2004).

The values of the ARI have an upper bound of one, where one equals perfectly equal clustering, zero is equal to a random clustering and a value lower than zero equals a worse performance than random clustering.

4.4 Results

To evaluate the performance of the eleven clustering methods outlined in Section 3.4, we apply each technique to a set of 48 data sets simulated using a set of parameters as described in Section 4.2. For each dataset, we compute the ARI relative to the original clustering, generating a 48 × 11 matrix of ARI's. This process is repeated five times, and the average ARI for each dataset and method is obtained by taking the mean of the five ARI's per method. The ARI's are presented in Table 9 in Appendix A. Figure 1, 2, 3 and 4 show the ARI's plotted for each type of data coding and for all methods. Each column represents the type of data coding (raw, doubled and categorical respectively). Each row represents the number of active variables (3, 6 and 20 respectively). The x-axis of each of the subplots shows whether the analysis used has noise variables and shows the number of clusters used.

The number of dimensions used by PCA and CA is determined by Kaiser's rule (Kaiser, 1960). If there are no dimensions with an eigenvalue larger than 1, we determine the number of dimensions by picking the dimensions that together explain at least 80% of the variance (Section 3.2.3). For RKM and FKM we set the number of dimensions to $D = \min((K-1),(Q-1))$, which is equal to the maximum number of dimension able to be chosen using these methods (Timmerman et al., 2010). Furthermore, the number of clusters used is equal to the number of clusters used to simulate the data.

Observing the results shown in Figure 1, we can see that, in the case of low overlap and balanced data, the raw tandem analysis, FKM and PAM are outperformed by other methods in all of the cases. However, this difference between all the methods is larger for FKM and PAM than for the tandem analysis. We can also see that when using raw or doubled data, K-means always performs best in the case of no noise variables being present. This also holds true for CCA in the case of the data being categorical. When noise variables are introduced, we can see that in some of the cases, RKM tends to perform better than standard K-means. When comparing raw data to doubled data, we can see that in the case of six active variables being used, raw tandem and RKM seem to benefit from this type of data coding. Comparing the raw and doubled data to categorically coded data, we can see that only in the case of three active variables being present, categorical data coding performs better.

Looking at Figure 2, we can see that using unbalanced data makes all methods perform worse compared to using balanced data. We can also that, in the case of raw and doubled data, K-means performs best when no noise is present and RKM performs best when noise is present. Next we see that categorical tandem outperforms CCA in all instances, implying that CCA suffers more from the data being unbalanced than categorical tandem.

Figure 3 shows the analyses applied on balanced data with a high overlap between different clusters. We can immediately see that the ARI's are much lower compared to the case of low overlap, which is to be expected. We also see that in no instance, a method applied on categorical data outperforms a method used on raw or doubled data. We can also see that RKM on doubled data performs best in the situation of noise variables being present, however, only in the situation of three active variables being present. In the case of six and twenty active variables, we see that K-means dominates in all situations.

Lastly, Figure 4 shows the methods applied on unbalanced data with a high overlap between the clusters. We can again see that RKM on doubled data outperforms K-means only in the situation of three active variables being used. In all the other cases, K-means out performs all methods.

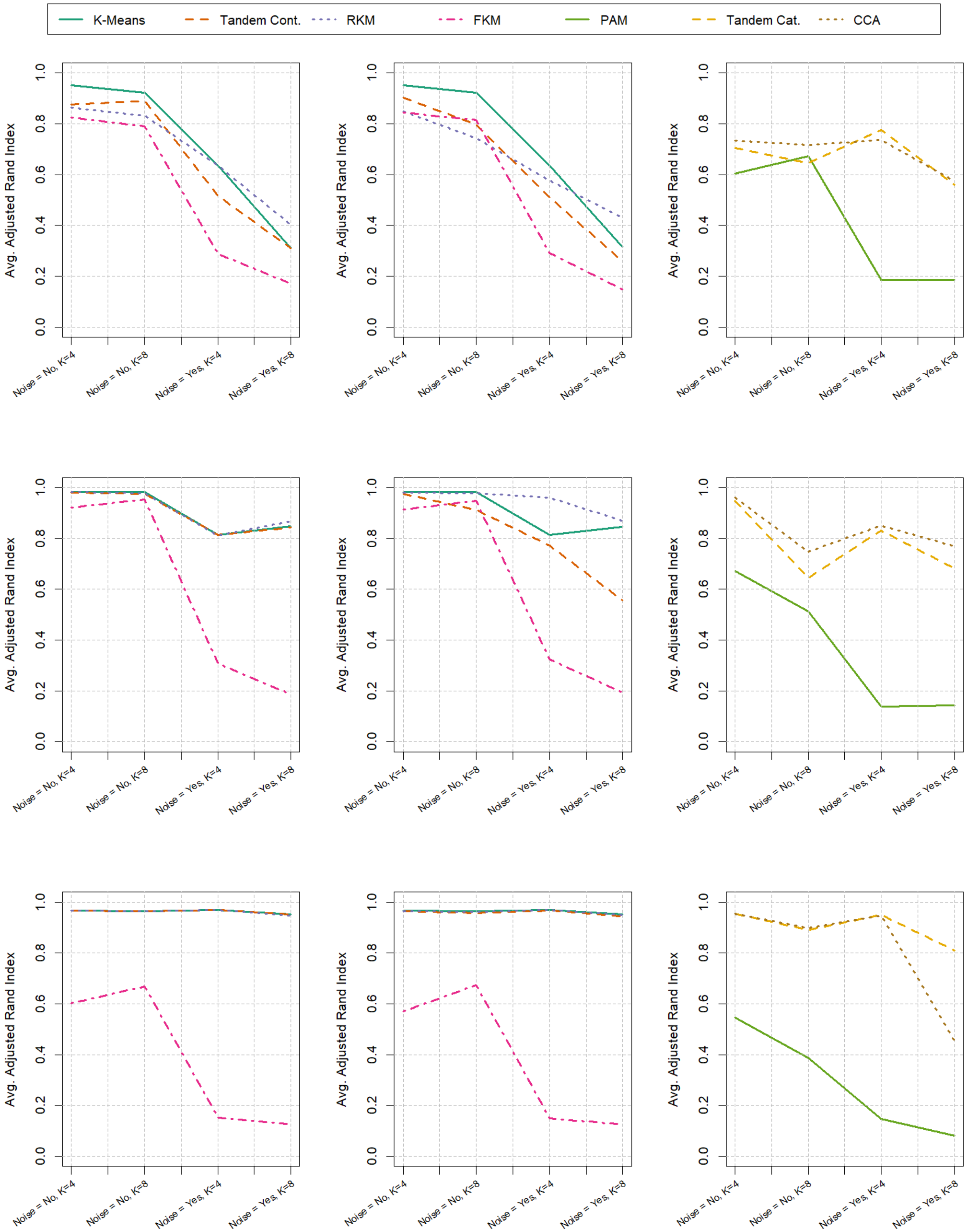

Figure 1: Plot of the average adjusted rand indices of eleven clustering methods with balanced data and low overlap using three (first row), six (second row) and twenty (third row) questions and applied on raw (first column), doubled (second column) and categorical (third column) data.

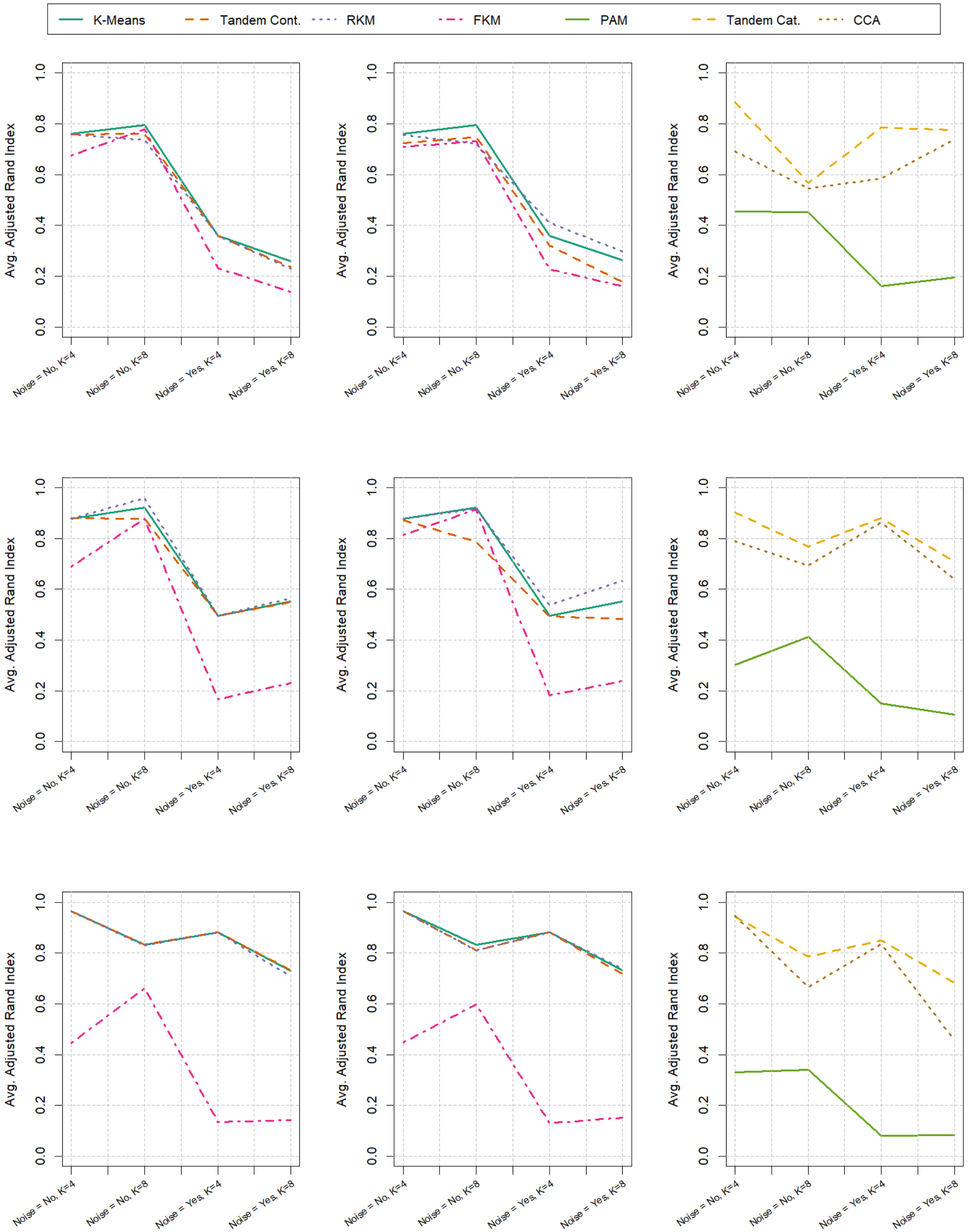

Figure 2: Plot of the average adjusted rand indices of eleven clustering methods with unbalanced data and low overlap using three (first row), six (second row) and twenty (third row) questions and applied on raw (first column), doubled (second column) and categorical (third column) data.

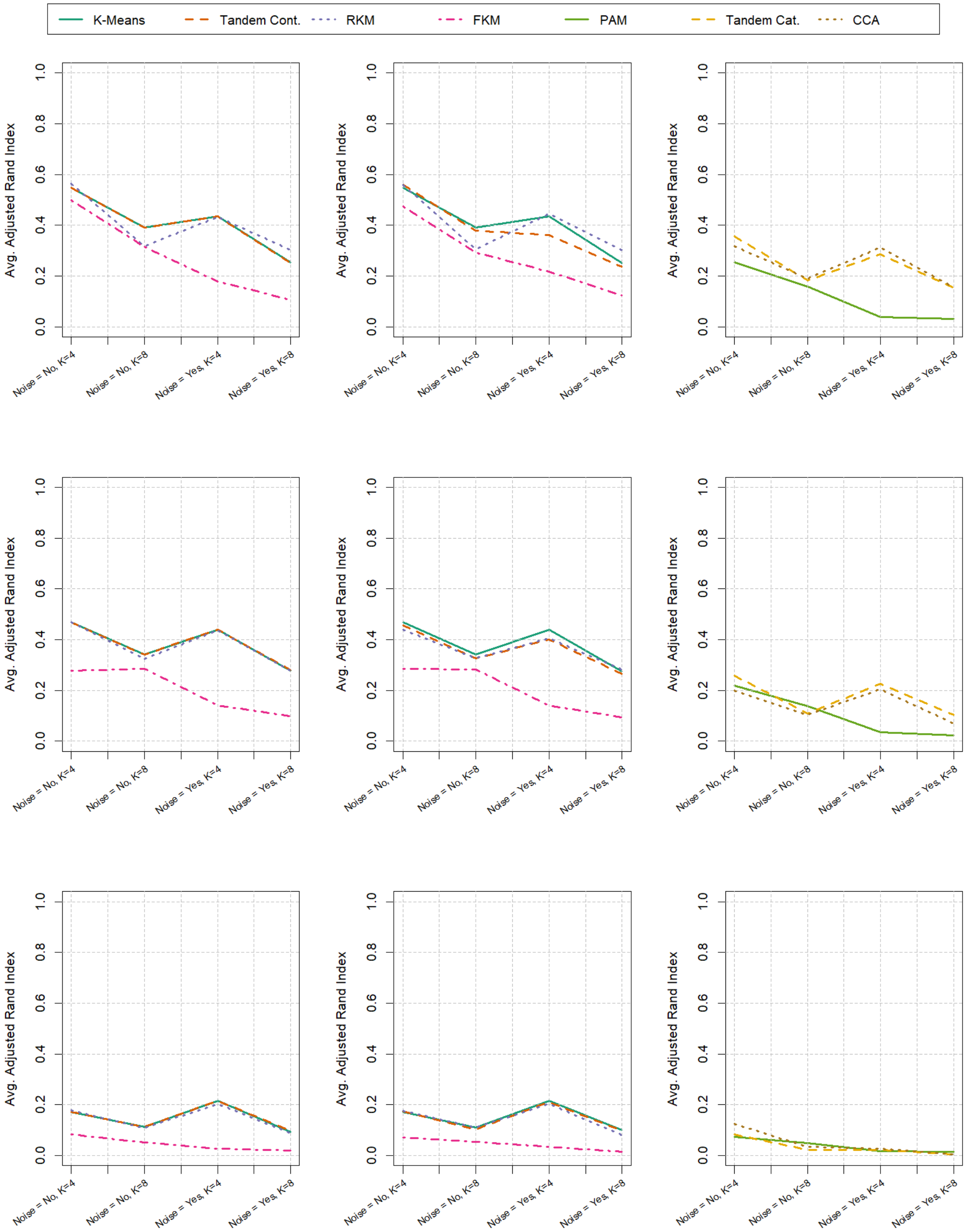

Figure 3: Plot of the average adjusted rand indices of eleven clustering methods with balanced data and high overlap using three (first row), six (second row) and twenty (third row) questions and applied on raw (first column), doubled (second column) and categorical (third column) data.

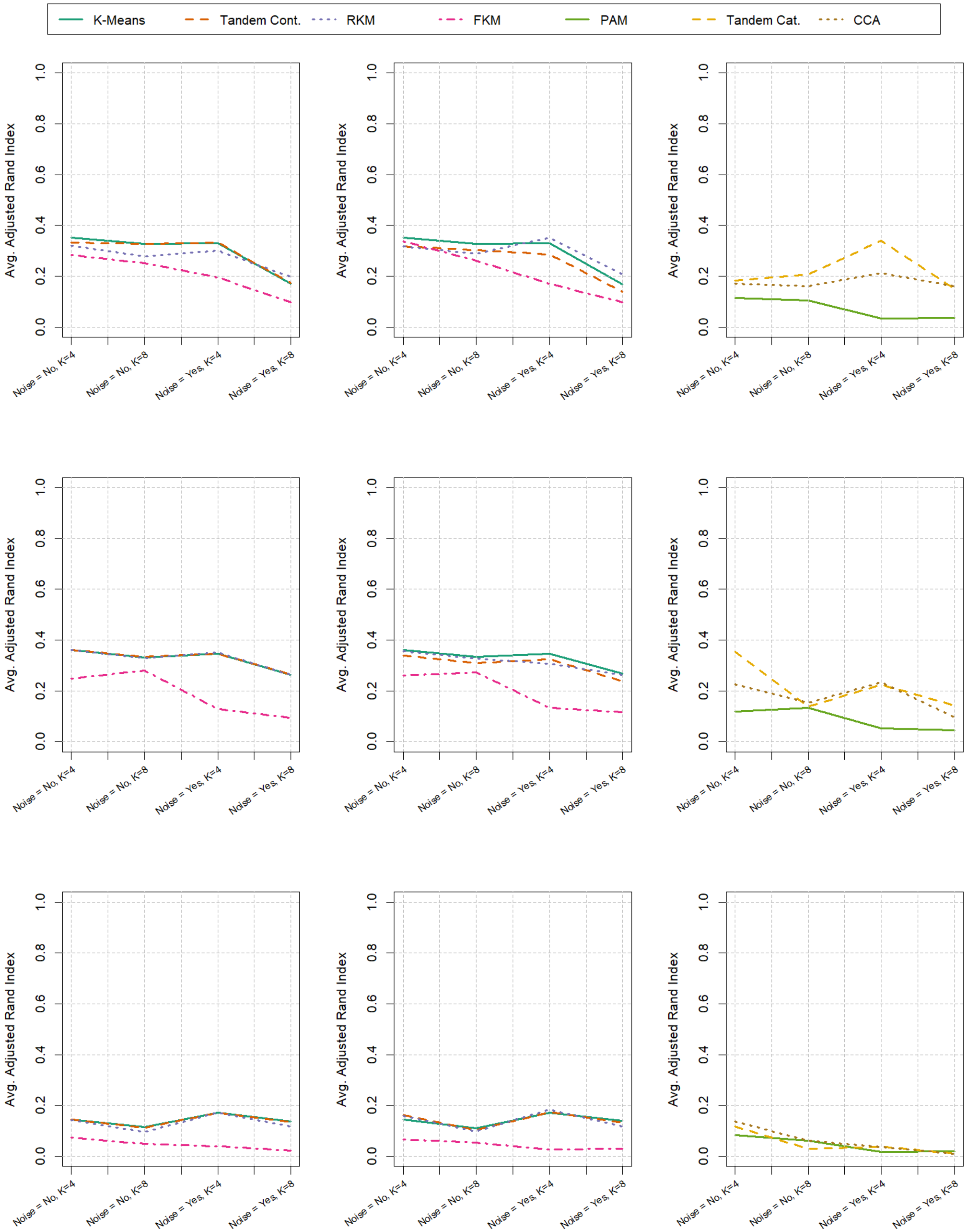

Figure 4: Plot of the average adjusted rand indices of eleven clustering methods with unbalanced data and high overlap using three (first row), six (second row) and twenty (third row) questions and applied on raw (first column), doubled (second column) and categorical (third column) data.

4.5 Conclusion of simulation study

The research consistently indicates that, in the absence of noise variables, the standard K-means clustering algorithm performs most effectively. This is expected, as the alternative methods incorporate some form of dimensionality reduction, which can result in the loss of information relevant to the underlying clustering structure if applied to data sets with exclusively active variables. In other words, the use of dimensionality reduction techniques on data sets with exclusively active variables may not fully capture the complexity of the clustering structure due to the exclusion of variables that may be meaningful for clustering observations.

Additionally, the research indicates that clustering the full data set is not optimal when noise variables are present. However, we do observe that the standard K-means algorithm performs comparably well in certain scenarios where there are 20 active and 60 noise variables. In the majority of cases with noise variables present, the reduced K-means method of doubled data and tandem analysis of categorical data emerge as the most effective approaches. The latter technique, in particular, tends to outperform the other methods when the dataset exhibits unbalanced cluster sizing.

The results also suggest that an increase in the overlap between cluster distributions leads to a corresponding decrease in the adjusted rand index. This outcome is anticipated, as greater overlap between clusters makes it more challenging to differentiate between them. In other words, the greater the overlap between clusters, the more difficult it becomes to accurately assign observations to their respective cluster.

5 Application

To show how the different methods work in practice, the previously discussed methods are applied on a real-world data set. The data set used is from world value surveys (Haerpfer et al., 2022). The surveys conducted by the scientists of world value surveys (WVS) focus mainly on the participants changing ethical values and their impact on social and political life. The surveys were conducted in almost 100 countries, but in this application the data is filtered to only contain a small group of countries. The data set used is the seventh wave of surveys, which took place from 2017 till 2022.

Table 2 shows the questions used in this analysis. Nineteen statements regarding ethical values and norms are given. Each statement is rated on a 10 point-scale. Here 1 equals never justifiable and 10 equals always justifiable. Furthermore, Table 3 shows variables D1-D4. These variables are not used to apply cluster analysis on, but rather as descriptive variables to compare the different clusters.

The subset of countries is chosen by comparing the countries based on the first four cultural dimensions of Hofstede (Hofstede, 1980). These four variables are: individualism, power distance, masculinity and uncertainty avoidance. Comparing these variables across different countries can give a rough sketch of the difference in culture between countries. Countries with similar scores

Table 2: WVS questionnaire. Each statement is rated on a scale from 1 (never justifiable) to 10 (always justifiable).

Label	Statement	Original Variable
Q1	Claiming government benefits to which you are not entitled	Q177
Q2	Avoiding a fare on public transport	Q178
Q3	Stealing property	Q179
Q4	Cheating on taxes if you have a chance	Q180
Q5	Someone accepting a bribe in the course of their duties	Q181
Q6	Homosexuality	Q182
Q7	Prostitution	Q183
Q8	Abortion	Q184
Q9	Divorce	Q185
Q10	Sex before marriage	Q186
Q11	Suicide	Q187
Q12	Euthanasia	Q188
Q13	For a man to beat his wife	Q189
Q14	Parents beating children	Q190
Q15	Violence against other people	Q191
Q16	Terrorism as a political, ideological or religious man	Q192
Q17	Having casual sex	Q193
Q18	Political violence	Q194
Q19	Death penalty	Q195

on these cultural dimensions can be grouped together, resulting in five clusters of countries with similar scores on the cultural dimensions. These clusters are: an Anglo cluster, an Nordic cluster, an Germanic cluster, an East-Asian cluster and an independent cluster (Hennig-Thurau et al., 2005). In our analysis, we choose a country from each of these five clusters, which results in the countries being used in the analysis being: the United States (Anglo), the Netherlands (Nordic), Germany (Germanic), China (East-Asian) and Japan (independent). Hofstede found two additional variables to add onto the cultural dimensions, however, these were not yet used in the time Hennig-Thurau et al. (2005) conducted their research, so these are not used to create the subset. For each country, a subset of 300 observations is randomly drawn, with the only condition being that the observation has a valid value for each statement and descriptive variable. This results in 1,500 observations

Table 3: Descriptive variables

Label	Variable	Original Variable	Categories
D1	Countries	B_COUNTRY_ALPHA	1. China 2. Germany 3. Japan 4. Netherlands 5. USA
D2	Age	X003R2	1. 16-29 years 2. 30-49 years 3. 50+ years
D3	Highest educational level	Q275R	1. Lower 2. Middle 3. Higher
D4	Income level	Q288R	1. Low 2. Medium 3. High

being used in the analysis.

In the application, we apply doubled RKM on the data set, as the simulation study showed this method to perform relatively well in most situations. For the doubled RKM analysis, we chose to include four dimensions and five clusters. The four dimensions are chosen following the simulation study, where $D = \min((K-1),(Q-1))$. The number of clusters are determined using the elbow plot in Figure 5, which results in five clusters being used.

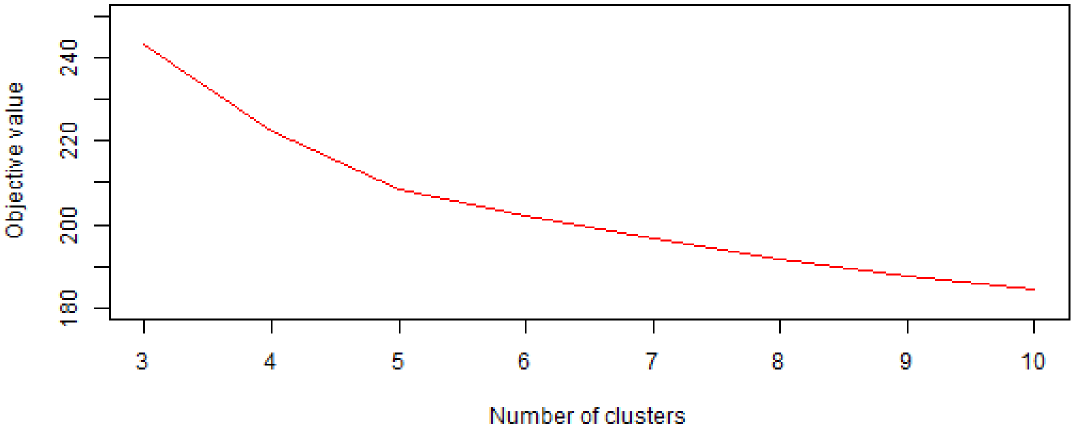

Figure 5: Elbow plot of double RKM

When observing Table 4, we can see which linear combinations of statements are chosen to represent each dimension of the RKM analysis. These values are called factor loadings.

First, when looking at the first dimension, we see that this dimension is mainly defined by questions 6, 7, 8, 9, 10, and 17. All these variables have a positive relation with the first dimension. This implies that the observations that have a high value in the first dimension of the reduced space believe that homosexuality, prostitution, abortion, divorce, sex before marriage and having casual sex should be justifiable.

The second dimension is mostly related with questions 3, 13, 15, 16 and 18. These five question all have a positive relation to this dimension. A high score in the second dimension indicates that the respondent believes stealing, a man beating his wife, violence, terrorism and political violence should be justifiable.

The third dimension is defined strongest by questions 2, 13 and 16. Question 2 has a negative relation to the third dimension, meaning that a high score in the third dimension indicates that this respondent does not think that avoiding a fare on public transport should be justifiable. Questions 13 and 16 have a positive relation to this dimension, indicating that respondents scoring high in this dimension think that a man beating his wife and terrorism should be justifiable.

Lastly, the fourth dimension is strongly related to questions 4, 11, 16 and 19, with question 4 and 19 being positively related and questions 11 and 16 being negatively related. This implies that respondents with a high score in the fourth dimension of the reduced space do think that cheating on taxes and the death penalty should be justifiable and suicide and terrorism should not.

Table 5 shows the doubled factor loadings. These loadings are calculated using Equations (14) and (15) on page 18. However, the values of d_{j_+} and d_{j_-} correspond to the distance of points to the origin in the correspondence analysis of ratings (CAr). To get the factor loadings in the RKM model, we calculate the doubled RKM factor loadings as $d_{j_-,RKM} = d_{j_+,RKM} \cdot \frac{d_{j_-}}{d_{j_+}}$. Here $d_{j_+,RKM}$ is the Euclidean distance of the RKM factor loading of question j to the origin and d_{j_+} and d_{j_-} are the Euclidean distances calculated using Equation (14) and (15).

Table 4: Doubled RKM Dimensions

Question	Dim 1	Dim 2	Dim 3	Dim 4
1	-0.078	0.196	-0.229	0.169
2	0.085	0.247	**-0.319**	0.288
3	0.050	**0.335**	-0.099	0.167
4	0.086	0.289	-0.268	**0.307**
5	0.033	0.283	-0.019	0.185
6	**0.430**	-0.072	0.278	0.245
7	**0.341**	0.037	-0.294	-0.254
8	**0.353**	-0.032	-0.019	-0.121
9	**0.316**	-0.065	0.174	0.146
10	**0.375**	-0.082	0.231	0.287
11	0.286	0.009	-0.145	**-0.327**
12	0.284	-0.078	0.107	0.028
13	0.009	**0.373**	**0.371**	-0.274
14	-0.071	0.255	-0.059	-0.170
15	0.041	**0.327**	0.024	-0.015
16	0.030	**0.427**	**0.420**	**-0.303**
17	**0.366**	0.065	-0.300	-0.217
18	0.035	**0.318**	0.024	0.124
19	-0.062	0.041	0.274	**0.352**

Table 5: Doubled RKM Dimensions

Question	Dim 1	Dim 2	Dim 3	Dim 4
1	0.012	-0.029	0.034	-0.025
2	-0.010	-0.030	0.038	-0.035
3	-0.002	-0.013	0.004	-0.006
4	-0.007	-0.022	0.021	-0.023
5	-0.002	-0.016	0.001	-0.011
6	**-0.643**	0.108	**-0.417**	**-0.367**
7	-0.144	-0.016	0.124	0.107
8	-0.283	0.026	0.016	0.097
9	**-0.495**	0.101	-0.273	-0.228
10	**-0.729**	0.159	**-0.448**	**-0.558**
11	-0.106	-0.003	0.054	0.122
12	**-0.356**	0.098	-0.134	-0.035
13	0.000	-0.012	-0.012	0.009
14	0.008	-0.031	0.007	0.020
15	-0.003	-0.026	-0.002	0.001
16	-0.001	-0.011	-0.011	0.008
17	-0.217	-0.039	0.177	0.128
18	-0.002	-0.021	-0.002	-0.008
19	0.045	-0.030	-0.200	-0.257

In Table 6, the four-dimensional coordinates of the cluster means are shown. These coordinates can be used in combination with the factor loadings in Table 4 to show how the clusters are formed.

Table 6: Cluster means

Cluster	Dim 1	Dim 2	Dim 3	Dim 4
1	-0.028	-0.111	0.106	0.062
2	0.437	-0.091	-0.039	-0.065
3	-0.481	-0.002	-0.058	-0.050
4	0.169	0.446	-0.166	0.156
5	0.128	1.677	0.414	-0.242

We can see that the first cluster has a relatively strong relation with the second and third dimension, where its relation with the second dimension is negative, whereas with the third dimension, the relation is positive. This indicates that people in the first cluster have a give a relatively low rating to question 2, 3, 15 and 18. This is confirmed by the average ratings per cluster shown in Table 7. Note that question 13 and 16 both have similar factor loadings in both dimensions and the mean of cluster one has similar scores for these dimensions, only in opposite directions. Making assumptions on the average ratings of these questions based on the cluster mean and factor loadings is therefore not advised.

The second cluster shows to have a high score in the first dimension. This indicates that respondents from the second cluster tend to give a high rating to questions 6, 7, 8, 9, 10 and 17, which is again confirmed by the average ratings in Table 7.

Respondents in the third cluster have a strong negative score in the first dimension, indicating that these respondents gave question 6, 7, 8, 9, 10 and 17 a low rating score.

The fourth cluster has a relatively high score in the second dimension. This shows us that respondents in this cluster gave high ratings to questions 3, 13, 15 and 16. The positive correlation with the first and fourth dimension shows us that respondents in this cluster also gave a high rating to questions 4, 6, 7, 8, 9, 10, 17 and 19 and gave low ratings to question 11. The negative score with the third dimension shows us that respondents in this cluster also rated question 2 high. Note that the factor loadings of the third dimension and the cluster means hint at the ratings for questions 13 and 16 being low. However, as the cluster mean has a higher score in the second dimension, the effect of this dimension is higher.

Lastly, the fifth cluster has a relative high score in the second dimension, similar to the fourth dimension, indicating a high rating for questions 3, 13, 15 and 16. The positive value in the first and third dimension show a high rating being given for questions 6, 7, 8, 9, 10 and 17 and a low rating being give to question 2. The fourth dimensions shows that respondents in this cluster rated questions 4 and 19 low and questions 11 and 16 high.

Table 7: Average ratings

Cluster	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
1	1.7	1.4	1.0	1.2	1.3	7.2	2.6	4.7	7.0	7.7	2.7	6.3	1.1	1.4	1.3	1.1	3.1	1.3	5.7
2	1.5	2.1	1.2	1.7	1.3	9.6	6.8	8.2	8.9	9.6	6.1	8.4	1.1	1.5	1.5	1.1	7.8	1.3	3.6
3	2.7	1.5	1.1	1.2	1.3	1.9	1.3	1.8	3.2	3.1	1.5	3.2	1.2	2.5	1.4	1.1	1.5	1.3	4.7
4	4.0	4.7	2.8	4.0	2.9	7.0	5.2	5.6	6.8	7.6	3.9	6.0	1.7	3.0	3.1	1.7	6.4	3.1	5.1
5	4.6	4.7	4.9	4.8	5.3	6.2	4.5	5.2	5.9	5.8	4.7	4.9	7.0	7.2	7.4	7.1	6.0	6.4	6.5

Finally, we investigated the adjusted rand indices (ARI) of all clustering results against each other, which are presented in Table 8. It should be noted that the actual clustering structure is unknown, making it impossible to compare the ARI's with the true clustering structure. However, assuming that the doubled RKM method performs relatively well, as suggested by the simulation study, some conclusions can be drawn.

The results reveal that the clustering obtained from the doubled RKM method does not show a high similarity with any of the methods. The simulation study showed that this was often the case when there were noise variables present. This finding suggests that this application data set may contain noise variables, and therefore, not all variables may be crucial for the underlying clustering structure.

Table 8: Application ARI

		Raw Data				Doubled Data				Categorical Data		
		KM	TD	RKM	FKM	KM	TD	RKM	FKM	PAM	TD	CCA
Raw Data	KM	1										
	TD	0.997	1									
	RKM	0.998	0.994	1								
	FKM	0.864	0.866	0.862	1							
Doubled Data	KM	1	0.997	0.998	0.864	1						
	TD	0.438	0.439	0.438	0.430	0.438	1					
	RKM	0.620	0.622	0.620	0.617	0.620	0.434	1				
	FKM	0.576	0.578	0.577	0.589	0.576	0.439	0.859	1			
Categorical Data	PAM	0.321	0.323	0.320	0.317	0.321	0.309	0.306	0.305	1		
	TD	0.141	0.141	0.141	0.146	0.141	0.133	0.192	0.185	0.118	1	
	CCA	0.344	0.346	0.345	0.354	0.344	0.325	0.437	0.432	0.285	0.292	1

Additionally, descriptive variables were added to the data set to show the composition of the different clusters based on the nationality, age, education level and income level of the respondents.

Figure 6 shows the descriptive values of the first cluster. Here we can see that this cluster consists of approximately 500 respondents with most respondents having a Japanese or German nationality. Most respondents are aged 50 years or older and have a middle to high educational level. The respondents in this cluster have a low to medium income.

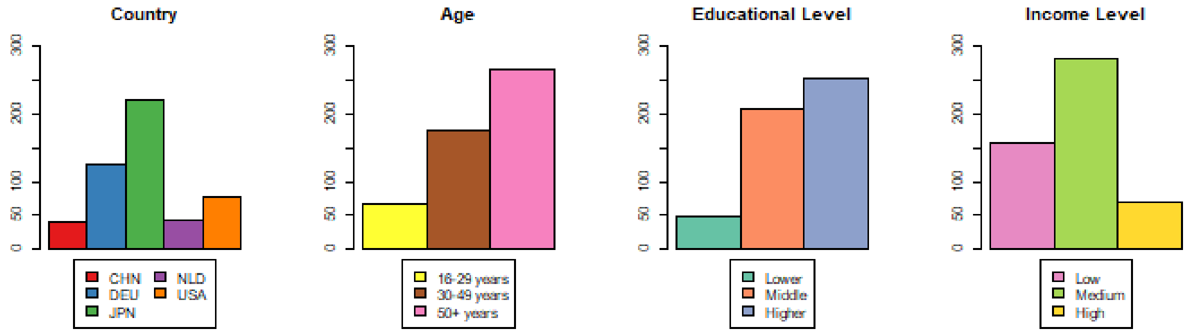

Figure 6: Bar Plots of descriptive variables from respondents in the first cluster

In Figure 7, we see the descriptive values of the second cluster, which consists of around 400 respondents. Here most respondents originate from Germany or the Netherlands and are mostly aged 50 years or older. Similar to the first cluster, respondents in this cluster have a middle to high educational level. The income level in this cluster is dominated by respondents with a medium income.

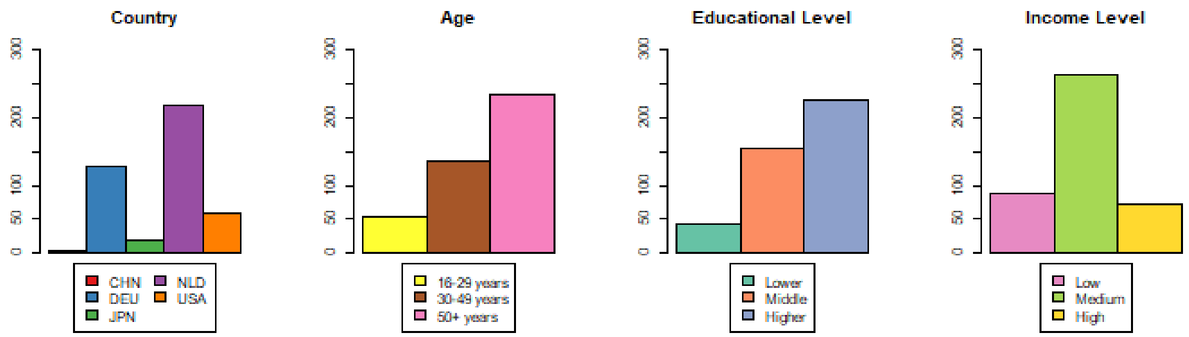

Figure 7: Bar Plots of descriptive variables from respondents in the second cluster

The descriptive values of the third cluster are shown in Figure 8. We can see that this cluster consists of around 400 respondents as well. The respondents mainly originate from China, with most people being aged 50 years or older. The educational level in this cluster is distributed fairly even across the lower, middle and higher level. Respondents in this cluster mostly have a low to medium income.

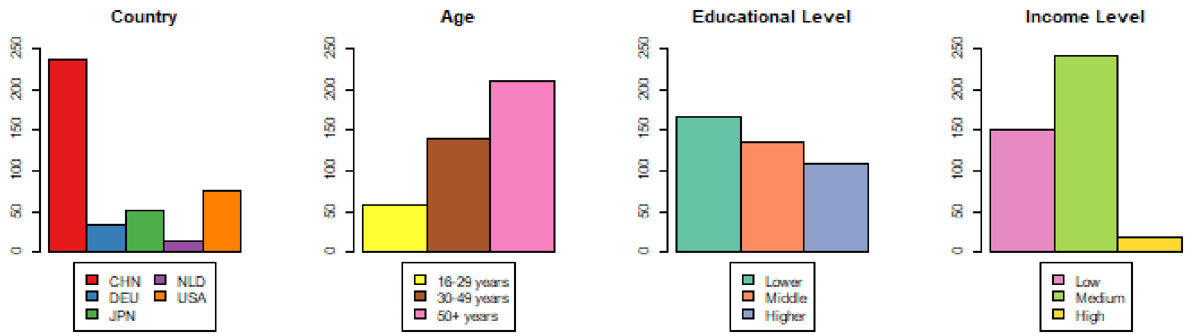

Figure 8: Bar Plots of descriptive variables from respondents in the third cluster

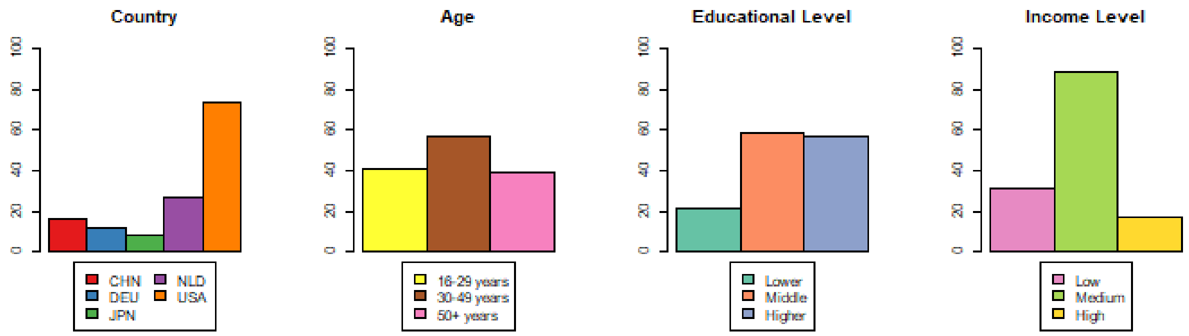

Figure 9: Bar Plots of descriptive variables from respondents in the fourth cluster

Figure 9 shows us the bar plot for the descriptive values of the fourth cluster, consisting of approximately 150 respondents. This cluster consists mainly of respondents from the USA. The age is distributed fairly even across the three categories: 16 to 29 years old, 30 to 49 years old and 50 years or older. This cluster mainly consists of respondents with a middle to high eductional level and a medium income.

The descriptive values of the fifth and final cluster are shown in Figure 10. This cluster contains around 20 respondent, most of American heritage. Similar to the fourth cluster, the age distribution is reasonably even as well. However, this cluster is mainly formed by respondents with a middle educational level combined with a medium income.

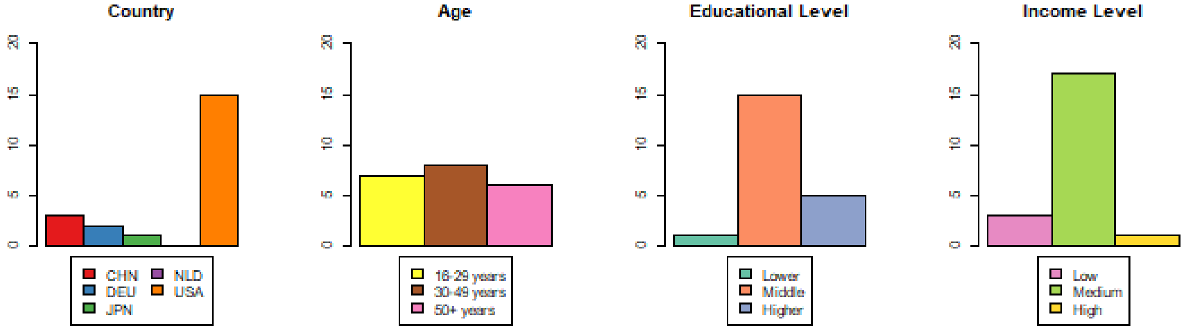

Figure 10: Bar Plots of descriptive variables from respondents in the fifth cluster

One additional application of the techniques discussed concerns a combination of doubled RKM with correspondence analysis of ratings (CAr), which was originally introduced by Greenacre (1984). A notable advantage of CAr is its ability to exhibit the spread of the variables. Specifically, Greenacre demonstrated that by doubling the rating matrix, a symmetry is established between the two poles of the bipolar rating variables. When the CA results of the doubled rating matrix are plotted, it displays pairs of column coordinates (representing negative and positive associations) that lie in a straight line passing trough the origin. The length of these lines and the proportion of the negative and positive sections of the lines offer valuable information regarding the distribution of the ratings. However, when applying doubled RKM, we do not apply RKM on the doubled rating matrix, but on the standardised data set, following Equation 17. Section 3.2.5 and specifically Equations (14) and (15) show us the relation between the doubled points in CAr. This method is also applied on the ratings of doubled RKM, to retrieve the doubled points. The depicted plot in Figure 11 illustrates this concept. The greater the distance between the poles, the higher the variance in opinion regarding that particular question. If a question has a higher positive association than negative, such as in the case of Q16, the pole that is located farthest from the origin exhibits the least association. Consequently, for Q16, this indicates that the negative association with this statement surpasses the positive association, implying that most people rated the sixteenth statement low. Thus, employing doubled RKM and computing the doubled points can provide meaningful insights into the distribution of the various rating variables.

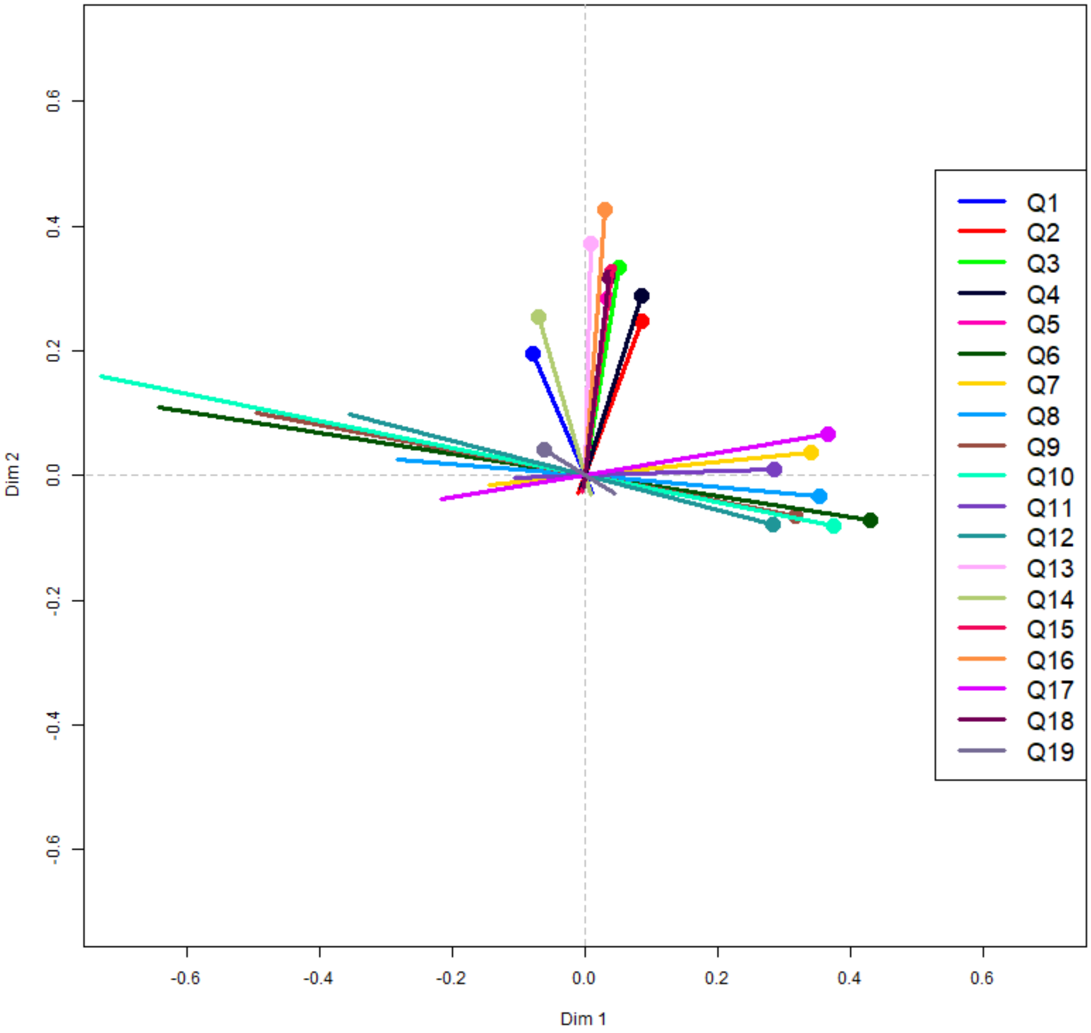

Figure 11: Doubled RKM map. The dotted end of the lines show the part of the line that has a positive association

6 Conclusion and Future Research

In this research, the main goal was answering the following research question:

> *How do different methods of cluster analyses combined with dimension reduction perform when applied on rating data?*

In order to address the research question at hand, we conducted a research utilising eleven distinct clustering methods. These methods used various data coding techniques and analytical approaches. Specifically, we utilised raw rating data, doubled rating data and categorical data as the three types of data coding. Through research done previously, we determined that doubling the rating data set in conjunction with simultaneous dimension reduction and cluster analysis could lead to more accurate clustering.

We conducted different types of analyses based on each of the three types of data coding. Firstly, we applied clustering analyses on the raw data set. That is, without any dimension reduction techniques used. For raw and doubled data, this involved regular K-means, while for categorical data, we employed the partitioning around medoids (PAM) algorithm. Secondly, we performed cluster analyses on the doubled rating data using a tandem approach. This involved applying dimension reduction prior to clustering the data. The objective of this was to assess the effect of noise variables on the clustering accuracy, as dimensions reduction should mitigate the influence of such variables on the analysis. This tandem analysis entailed principal component analysis (PCA) followed by K-means for raw rating data, correspondence analysis of ratings (CAr) for doubled rating data, and multiple correspondence analysis (MCA) followed by K-means for categorical data. Lastly, we employed methods that concurrently reduced the number of dimensions of the data set while clustering the data. Past research has suggested that this approach could, in certain cases, lead to a more accurate clustering structure than the tandem approach. The methods utilised in this regard included reduced K-means (RKM) and factorial K-means (FKM) for raw and doubled data, and cluster correspondence analysis (CCA) for categorical data.

To determine which of these methods performed best, we conducted a simulation study. In this study, we simulated rating scale data by drawing variables from a Gaussian mixture model and discretising the values. This allowed us to exercise control over certain parameters, such as the degree of overlap between variables or the number of clusters. The eleven clustering methods under investigation were then applied on these simulated data sets, and the results were assessed by comparing the original cluster structure with the clustering obtained from the different methods. We used the adjusted rand index (ARI) to evaluate the performance of each method. This metric produces a score with an upper bound of one, with a higher score indicating greater similarity between the clustering produced by a particular method and the original clustering structure.

The simulation study revealed that in instances where no noise variables were present, clustering on the complete data set proved to be the most effective approach for recovering the original clustering structure. This finding was not surprising, as implementing dimension reduction methods on a data set lacking noise variables would result in a loss of information on the original clustering structure. However, when assessing methods utilised on data sets containing noise variables, we observed that the doubled RKM method demonstrated the highest degree of efficacy across the majority of cases.

To conclude this research, we utilised the best performing method identified in the simulation study, namely the doubled RKM approach, to analyse a real-world data set. This data set comprised opinions on particular ethical statements from five different countries. Our aim was to demonstrate how this method can be employed to conduct marketing research and illustrate its ability to visualise the outcomes of cluster analysis. Furthermore, we showed the application of the doubled RKM analysis and its ability to visualise the distribution of the rating variables.

Comparative studies pertaining to rating scale data which compare multiple clustering methods is pretty scarce. Moreover, while Greenacre introduced the technique of doubling rating data in 1984, it has yet to be applied in conjunction with this particular combination of methods. This lack of research done in the literature underscores the significance of our research to the marketing field.

In this study, there is room for future research to expand upon the methods employed. Specifically, the rating data could be subjected to more analyses, such as MCA K-means or GROUPALS (Hwang et al., 2006; Van Buuren & Heiser, 1989). These methods can be compared with the techniques utilised in this research. In the Literature Review, we discussed rank order and successive categories data. These types of data require distinct clustering techniques, unlike those used in this research. By comparing the performance of methods applied on these types of data against those utilised in this research, insights could be gained into the effectiveness of cluster analyses in varying conditions.

The simulation study could be broadened to include a wider range of parameters. Due to the computational limitations, we opted for a specific set of parameters in this research. However, with additional time or computing power, more varied parameter values could be examined to provide further insight into the optimal methods for specific cases. For instance, the number of variables and noise variables could be increased beyond the three levels studied here. This study considered only no noise being present or equalled the noise variables to three times the number of active variables. These values represent extreme cases, and a wider range of noise levels would be more informative. Additionally, the degree of overlap between Gaussian densities could be further diversified. The assumption of the ratings being normally distributed does not always hold true in practice, thus in future research, it could be interesting to investigate the accuracy of clustering when the ratings are drawn from different distributions, such as a multinomial distribution. Finally, although the correlation between questions within the same cluster was not controlled in this research, examining these effect of varying correlation values could provide further understanding of the performance of the techniques discussed.

References

Bellman, R., & Kalaba, R. E. (1959, 11). On adaptive control processes. *IRE transactions on automatic control*, *4*(2), 1–9. doi: 10.1109/tac.1959.1104847

Benzécri, J. (1973). *L'analyse des données. volume ii. l'analyse des correspondances.* Paris, France: Dunod.

Chen, L. (2009). Curse of dimensionality. In *Encyclopedia of database systems* (pp. 545–546). Boston, MA: Springer US. doi: 10.1007/978-0-387-39940-9_133

Dave, R. N. (1991, 11). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, *12*(11), 657–664. doi: 10.1016/0167-8655(91)90002-4

De Leeuw, J., Young, F. W., & Takane, Y. (1976, 12). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, *41*(4), 471–503. doi: 10.1007/bf02296971

De Soete, G., & Carroll, J. D. (1994). *K-means clustering in a low-dimensional Euclidean space.* Springer Berlin Heidelberg. doi: 10.1007/978-3-642-51175-2_24

Fisher, R. (1940, 1). THE PRECISION OF DISCRIMINANT FUNCTIONS. *Annals of eugenics*, *10*(1), 422–429. doi: 10.1111/j.1469-1809.1940.tb02264.x

Fränti, P., & Sieranoja, S. (2018, 12). K-means properties on six clustering benchmark datasets. *Applied Intelligence*, *48*(12), 4743–4759. doi: 10.1007/s10489-018-1238-7

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* London, UK: Academic Press.

Guttman, L. (1946, 6). An Approach for Quantifying Paired Comparisons and Rank Order. *Annals of Mathematical Statistics*, *17*(2), 144–163. doi: 10.1214/aoms/1177730977

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., … Puranen, B. (2022). *World values survey: Round seven - country-pooled datafile version 5.0.* Madrid, Spain Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi: 10.14281/18241.20

Harpe, S. E. (2015, 11). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*(6), 836–850. doi: 10.1016/j.cptl.2015.08.001

Hennig-Thurau, T., Gwinner, K. P., Gremler, D. D., & Paul, M. J. (2005). *Managing Service Relationships in a Global Economy: Exploring the Impact of National Culture on the Relevance of Customer Relational Benefits for Gaining Loyal Customers.* Emerald Publishing Limited. doi: 10.1016/s1474-7979(04)15002-3

Hofstede, G. (1980). *Culture's Consequences.* SAGE Publications, Incorporated.

Hsu, T.-C., & Feldt, L. S. (1969, 11). The Effect of Limitations on the Number of Criterion Score Values on the Significance Level of the F-Test. *American Educational Research Journal*, *6*(4), 515. doi: 10.2307/1162248

Hubert, L., & Arabie, P. (1985, 12). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. doi: 10.1007/bf01908075

Hwang, H., Dillon, W. P., & Takane, Y. (2006, 3). An Extension of Multiple Correspondence Analysis for Identifying Heterogeneous Subgroups of Respondents. *Psychometrika*, *71*(1), 161–171. doi: 10.1007/s11336-004-1173-x

Jolliffe, I. T., & Cadima, J. (2016, 4). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, *374*(2065), 20150202. doi: 10.1098/rsta.2015.0202

Kaiser, H. F. (1960, 4). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. doi: 10.1177/001316446002000116

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* New York, USA: John Wiley & Sons Inc.

Kaushik, M., & Mathur, B. (2014). Comparative study of k-means and hierarchical clustering techniques. *International Journal of Software & Hardware Research in Engineering*, *2*(6), 93–98.

Kumar, P., & Wasan, S. (2011). Comparative study of k-means, pam and rough k-means algorithms using cancer datasets. *International Symposium on Computing, Communication, and Control*, *1*, 136–140.

MacQueen, J. (1967, 1). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, *1*(14), 281–297.

Madhulatha, T. S. (2011). *Comparison between K-Means and K-Medoids Clustering Algorithms.* Springer Science+Business Media. doi: 10.1007/978-3-642-22555-0_48

Maitra, R., & Melnykov, V. (2010, 1). Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms. *Journal of Computational and Graphical Statistics*, *19*(2), 354–376. doi: 10.1198/jcgs.2009.08054

Mathivanan, N. M. N., Ghani, N. A. M., & Janor, R. M. (2019). A comparative study on dimensionality reduction between principal component analysis and k-means clustering. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(2), 752–758. doi: 10.11591/ijeecs.v16.i2.pp752-758

Melnykov, V. (2016, 3). Merging Mixture Components for Clustering Through Pairwise Overlap. *Journal of Computational and Graphical Statistics*, *25*(1), 66–90. doi: 10.1080/10618600.2014.978007

Melnykov, V., Chen, W.-C., & Maitra, R. (2012, 11). MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms. *Journal of Statistical Software*, *51*(12). doi: 10.18637/jss.v051.i12

Myers, K., & Winters, N. C. (2002, 2). Ten-year review of rating scales. I: overview of scale functioning, psychometric properties, and selection. *Journal of the American Academy of Child and Adolescent Psychiatry*, *41*(2), 114–22. doi: 10.1097/00004583-200202000-00004

Pearson, K. (1901, 1). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. doi: 10.1080/14786440109462720

Steinley, D. (2004, 9). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods*, *9*(3), 386–396. doi: 10.1037/1082-989x.9.3.386

Thorndike, R. L. (1953, 12). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. doi: 10.1007/bf02289263

Timmerman, M. E., Ceulemans, E., Kiers, H. A., & Vichi, M. (2010, 7). Factorial and reduced K-means reconsidered. *Computational Statistics Data Analysis*, *54*(7), 1858–1871. doi: 10.1016/j.csda.2010.02.009

Torgerson, W. S. (1958). *Theory and methods of scaling.* Wiley.

Van Buuren, S., & Heiser, W. J. (1989, 1). Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, *54*(4), 699–706. doi: 10.1007/bf02296404

Van De Velden, M. (2004, 3). Optimal Scaling of Paired Comparison Data. *Journal of Classification*, *21*(1), 89–109. doi: 10.1007/s00357-004-0007-y

Van De Velden, M., D'Enza, A. I., & Palumbo, F. (2017, 3). Cluster Correspondence Analysis. *Psychometrika*, *82*(1), 158–185. doi: 10.1007/s11336-016-9514-0

Vichi, M., & Kiers, H. A. (2001, 7). Factorial k-means analysis for two-way data. *Computational Statistics Data Analysis*, *37*(1), 49–64. doi: 10.1016/s0167-9473(00)00064-5

Yamamoto, M., & Hwang, H. (2014, 1). A General Formulation of Cluster Analysis with Dimension Reduction and Subspace Separation. *Behaviormetrika*, *41*(1), 115–129. doi: 10.2333/bhmk.41.115

Appendix A: Table of Adjusted Rand Indices

Table 9: The average adjusted rand index, across all methods and data sets, relative to the original clustering structure , where the leftmost columns indicates the parameters used to generate each dataset. The green cells indicate the highest ARI of the corresponding data set.

Data Set					Raw Data				Doubled Data				Categorical Data		
Q	N	K	Avg. Overlap	Balanced	KM	TD	RKM	FKM	KM	TD	RKM	FKM	PAM	TD	CCA
3	0	4	0.001	true	0.952	0.876	0.864	0.826	0.952	0.904	0.850	0.844	0.604	0.704	0.735
3	0	4	0.001	false	0.760	0.758	0.758	0.675	0.760	0.725	0.755	0.709	0.455	0.887	0.693
3	0	4	0.100	true	0.548	0.548	0.565	0.500	0.548	0.560	0.559	0.475	0.254	0.358	0.318
3	0	4	0.100	false	0.351	0.332	0.321	0.283	0.351	0.317	0.317	0.337	0.114	0.182	0.171
3	0	8	0.001	true	0.924	0.890	0.831	0.791	0.924	0.797	0.743	0.816	0.673	0.645	0.718
3	0	8	0.001	false	0.796	0.760	0.737	0.778	0.796	0.749	0.721	0.731	0.453	0.568	0.545
3	0	8	0.100	true	0.391	0.391	0.318	0.314	0.391	0.379	0.306	0.292	0.158	0.182	0.189
3	0	8	0.100	false	0.327	0.327	0.279	0.252	0.327	0.303	0.287	0.261	0.104	0.208	0.160
3	9	4	0.001	true	0.637	0.517	0.636	0.287	0.637	0.510	0.577	0.291	0.185	0.776	0.737
3	9	4	0.001	false	0.359	0.359	0.359	0.233	0.359	0.320	0.411	0.228	0.161	0.786	0.584
3	9	4	0.100	true	0.434	0.434	0.433	0.178	0.434	0.361	0.446	0.217	0.039	0.285	0.313
3	9	4	0.100	false	0.331	0.332	0.301	0.194	0.331	0.286	0.351	0.170	0.033	0.339	0.212
3	9	8	0.001	true	0.310	0.311	0.404	0.171	0.314	0.260	0.431	0.148	0.185	0.559	0.576
3	9	8	0.001	false	0.258	0.237	0.229	0.138	0.263	0.179	0.297	0.161	0.194	0.776	0.739
3	9	8	0.100	true	0.253	0.251	0.302	0.106	0.251	0.236	0.302	0.124	0.032	0.153	0.155
3	9	8	0.100	false	0.171	0.174	0.198	0.096	0.169	0.138	0.207	0.096	0.036	0.152	0.162
6	0	4	0.001	true	0.983	0.980	0.983	0.921	0.983	0.977	0.981	0.915	0.672	0.948	0.961
6	0	4	0.001	false	0.878	0.880	0.878	0.688	0.878	0.873	0.879	0.814	0.301	0.902	0.789
6	0	4	0.100	true	0.468	0.468	0.468	0.278	0.468	0.456	0.439	0.286	0.218	0.257	0.199
6	0	4	0.100	false	0.360	0.360	0.360	0.247	0.360	0.338	0.357	0.260	0.118	0.353	0.226
6	0	8	0.001	true	0.983	0.977	0.979	0.954	0.983	0.912	0.979	0.950	0.513	0.646	0.747
6	0	8	0.001	false	0.921	0.877	0.958	0.879	0.921	0.786	0.918	0.914	0.412	0.767	0.694
6	0	8	0.100	true	0.342	0.342	0.324	0.286	0.342	0.327	0.327	0.282	0.139	0.109	0.104
6	0	8	0.100	false	0.332	0.333	0.330	0.279	0.333	0.309	0.327	0.272	0.134	0.138	0.152
6	18	4	0.001	true	0.813	0.813	0.811	0.307	0.813	0.772	0.962	0.325	0.138	0.831	0.850
6	18	4	0.001	false	0.495	0.495	0.495	0.167	0.495	0.493	0.538	0.183	0.149	0.880	0.864
6	18	4	0.100	true	0.440	0.440	0.437	0.141	0.440	0.400	0.404	0.141	0.034	0.225	0.204
6	18	4	0.100	false	0.346	0.346	0.350	0.128	0.346	0.325	0.308	0.134	0.053	0.223	0.233
6	18	8	0.001	true	0.848	0.843	0.869	0.188	0.847	0.557	0.870	0.195	0.143	0.681	0.770
6	18	8	0.001	false	0.551	0.550	0.564	0.230	0.551	0.483	0.632	0.239	0.107	0.708	0.640
6	18	8	0.100	true	0.278	0.280	0.278	0.099	0.274	0.266	0.283	0.093	0.023	0.103	0.070
6	18	8	0.100	false	0.264	0.266	0.264	0.093	0.267	0.239	0.264	0.117	0.045	0.144	0.097
20	0	4	0.001	true	0.967	0.967	0.967	0.603	0.967	0.966	0.966	0.572	0.547	0.955	0.955
20	0	4	0.001	false	0.966	0.966	0.966	0.446	0.966	0.966	0.966	0.449	0.331	0.943	0.947
20	0	4	0.100	true	0.172	0.172	0.178	0.082	0.172	0.174	0.176	0.072	0.073	0.083	0.124
20	0	4	0.100	false	0.144	0.144	0.143	0.074	0.144	0.161	0.160	0.065	0.082	0.117	0.138
20	0	8	0.001	true	0.966	0.966	0.966	0.669	0.966	0.959	0.963	0.674	0.387	0.891	0.900
20	0	8	0.001	false	0.832	0.832	0.830	0.662	0.832	0.812	0.812	0.597	0.341	0.787	0.666
20	0	8	0.100	true	0.113	0.113	0.108	0.051	0.111	0.102	0.111	0.054	0.050	0.023	0.034
20	0	8	0.100	false	0.116	0.113	0.096	0.049	0.111	0.104	0.098	0.053	0.061	0.029	0.061
20	60	4	0.001	true	0.969	0.969	0.969	0.151	0.969	0.968	0.970	0.149	0.147	0.951	0.948
20	60	4	0.001	false	0.883	0.883	0.881	0.134	0.883	0.881	0.882	0.129	0.081	0.850	0.836
20	60	4	0.100	true	0.216	0.216	0.204	0.026	0.216	0.211	0.204	0.035	0.017	0.023	0.026
20	60	4	0.100	false	0.172	0.172	0.171	0.040	0.172	0.174	0.183	0.027	0.016	0.036	0.036
20	60	8	0.001	true	0.953	0.953	0.949	0.126	0.953	0.946	0.951	0.126	0.081	0.810	0.452
20	60	8	0.001	false	0.731	0.732	0.708	0.141	0.732	0.718	0.740	0.153	0.083	0.683	0.455
20	60	8	0.100	true	0.092	0.098	0.089	0.020	0.101	0.099	0.081	0.014	0.015	0.004	0.005
20	60	8	0.100	false	0.138	0.135	0.118	0.022	0.139	0.131	0.117	0.029	0.019	0.012	0.010

Appendix B: Description of R Code

simulation.R

This code performs the simulation study. The data is drawn from a Gaussian mixture model with controlled overlap. Five data sets are created for every possible parameter combination and for every type of data coding. The eleven cluster methods are applied on each data set and the average adjusted rand indices are calculated relative to the original clustering structure.

application.R

This code performs the application. Data from WVS is used. The data is filtered to only contain the relevant variables. We sampled 300 observations from five different countries to create a dataset containing 1500 observations. The data is again coded in three different ways. The eleven clustering methods are performed on this data set. The ARI is calculated from the clustering of all methods compared to each other. Lastly, doubled RKM is performed on the data to show the versatility of the method and its visualisation. An elbow plot is made to determine the number of clusters. For each cluster, bar plots are made to show the descriptive variables and how the different clusters are composed.