Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis

Invariant Coordinate Selection as dimensionality reduction for cluster analysis

MSc Econometrics and Management Science

Programme: Business Analytics & Quantitative Marketing

Author:

Róża Trybulska 565257

Supervisor:

dr. Aurore Archimbaud

Second assessor:

drs. Jeffrey Durieux

August 2, 2022

Abstract

Clustering is a recognised unsupervised learning approach for data analysis. Recently, the amount of available data has increased exponentially and what corresponds with it, the dimensions of datasets. Big dimensionality is troublesome for the analysis and clustering methods used alone encounter problems. This paper proposes applying tandem clustering, namely performing dimension reduction first, before clustering. Cluster analysis is then done on a lower-dimensional subspace. Firstly, the well-known method called Principal Component Analysis (PCA) is used for dimensionality reduction. However, since it only focuses on the variability of the data and is sensitive to data transformations, a new method for multivariate analysis, namely Invariant Coordinate Selection (ICS), is considered. ICS depends on the joint diagonalization of two scatter matrices, and its properties of affine invariance and finding Fisher's linear discriminant subspace for mixture of elliptical distributions, are advantageous. The fact that ICS can find groups even without knowing its memberships, enables one to use it for clustering purposes. This paper aims to improve cluster identification by performing various tandem clusterings. It compares cluster validity indexes over the original clustering and also the clustering done on the subspaces obtained by the PCA and ICS methods. A simulation study shows improvement in cluster identification but not for very high dimensions. The performance of the proposed approach is also tested on a real dataset.

Keywords: Affine Invariance, Clustering, ICS, PCA, Scatter Matrices, Tandem Clustering

Contents

List of Abbreviations

ARI	Adjusted Rand Index
COV	Covariance Matrix
COV_4	Matrix of Fourth Moments
EM	Expectation Maximization
GMM	Gaussian Mixture Model
IC	Invariant Coordinate
ICS	Invariant Coordinate Selection
MCD	Minimum Covariance Determinant
PC	Principal Component
PCA	Principal Component Analysis

1 Introduction

Nowadays, the amount of available data brings many challenges. An efficient way of analysing data by getting insights into their structure groups is clustering. Clustering is a popular unsupervised learning method. Clusters are formed such that the points in one cluster are similar to each other or dissimilar to observations in other groups. For the datasets with large dimensions, it is known that some commonly used clustering algorithms (e.g., K-means) can underperform, for example, due to the local minimum problem (Ding et al., 2002). Different solutions and different clustering methods were proposed. One of the ideas is using dimension reduction first, ahead of clustering.

Shrinking the feature space is beneficial for many complex algorithms as it increases accuracy and efficiency (Anowar et al., 2021). The literature contains various examples of feature reduction done by Principal Component Analysis (PCA), often followed by the K-means clustering algorithm (Kaya et al. (2017), Lee et al. (2009)). This approach is called tandem clustering. It has been used frequently but also criticized (Soete and Carroll, 1994). When tandem clustering uses PCA as the dimension reduction step, it only considers the subspace explained by the variability, meaning that finding the structure of the data is limited. Therefore, replacing PCA with another method that can find structure of the information, is compelling. Among known dimension reduction methods, the Invariant Coordinate Selection (ICS) method can be of interest. Indeed, it is a modern technique for multivariate analysis, which has been proven useful for outlier identification and dimensionality reduction on different types of data (Archimbaud et al. (2018b), Fischer et al. (2017)). In fact, ICS focuses on finding the structure of the data and it considers the general kurtosis. Unlike PCA, the ICS method is affine invariant. This property indicates that output is not affected by scale or units when the data transformation is performed. Among other advantages, as shown by Tyler et al. (2009), in the context of mixture of elliptically symmetric distributions, the ICS method can recognise groups, even if the group membership is unknown. By maximizing the general kurtosis, it finds small clusters, that is useful, for example, for outlier identification. Minimizing the general kurtosis, on the other hand, finds large groups, which is valuable for clustering. However, this method has not yet been studied much in the context of cluster analysis. Therefore, we would like to use its property of identifying groups for clustering.

In this thesis, we aim to improve clustering by extending and evaluating the performance of ICS for tandem clustering. We want to investigate the idea of doing dimension reduction first, before

clustering, so that the cluster analysis is done on the low-dimensional subspace. We propose the following research question:

"Does the ICS dimension reduction before clustering improve cluster identification?"

We answer this question by comparing the performance of ICS and another dimension reduction method, namely PCA. The analysis is conducted on simulated data with three different dimension sizes as well as on a real dataset. After completing the reduction step, we proceed with cluster analysis. We perform and compare two clustering techniques, the K-means, and the Expectation-Maximization (EM) algorithm. We find that ICS dimension reduction can improve cluster identification but it depends on the data design. When the dimension is equal to 10, ICS does not encounter any problems and the clustering validation index achieves high values. However, it is not true if the data is split such that one group consists of around 20% of all observations. For higher dimensions, we obtain worse results.

This paper is organized as follows. In section 2, the literature on the subject is discussed. Section 3 presents the applied methodology, which consists of dimension reduction and clustering algorithms, as well as the evaluation methods. Section 4 describes the simulation design and results per scenario. In section 5, the description and the results of a real dataset are discussed. Lastly, the conclusion and discussion can be found in section 6.

2 Literature

We first review the existing literature concerning different clustering methods and their comparison. Kaya et al. (2017) define clustering as a process that finds complex data relationships, that are hidden in data sets, and presents groups of observations with similar characteristics. There are numerous clustering methods, however, K-means and the EM algorithms are mostly used (Kinnunen et al., 2011). K-means clustering is an unsupervised learning method that aims to divide data into K clusters where each point belongs to the group with the nearest mean (Kaya et al., 2017). However, Dempster et al. (1977) state that, unlike K-means, the EM algorithm is an example of a soft clustering, since clusters are formed based on the membership probabilities.

One of the difficulties with clustering, is to know which approach is best. To compare them, we can follow Kinnunen et al. (2011). They examine many clustering algorithms, including K-means and the EM. Their paper on text-independent speaker recognition suggests using the Detection

Error Tradeoff curve for choosing the right method. It is the compromise between a False Acceptance Rate and a False Rejection Rate. However, according to Shirkhorshidi et al. (2015), the Rand Index is the most common index for clustering validity. The Rand Index, introduced by Rand (1971) , measures similarity between two data groups. The improved version that is adjusted for the number and size of clusters, the Adjusted Rand Index (Hubert and Arabie, 1985), is frequently used. Another well-known measure for cluster validity and finding the right choice of groups is the average silhouette coefficient (Rousseeuw, 1987). It determines how well the data was grouped, and the results are shown in the silhouette graphs for visual analysis. Rousseeuw (1987) states that one can avoid clustering on the outliers by considering graphical displays. Additionally, Hennig (2022) emphasizes that there is no universal definition of a good cluster. As a consequence, he investigates many criteria of recovering true groups, one of which is the Adjusted Rand Index.

To improve clustering, especially for data with many dimensions, we research the idea of tandem clustering. Tandem clustering is a method that performs dimension reduction first, before clustering (Vichi and Kiers, 2001). Having a low-dimensional subspace is beneficial. Such representation of clusters enables easier identification of the groups, especially for two or three dimensions when the graphical analysis is possible (Soete and Carroll, 1994).

According to Ma and Yuan (2019), PCA is the most used dimension reduction method. It projects high-dimensional data into low-dimensional subspace by using linear transformation. Dimension reduction is done by keeping components that have the largest variance, which corresponds to the most amount of information (Hasan and Abdulazeez, 2021). However, Ma and Yuan (2019) claim that while decreasing dimensions enables computationally tractable and easier analysis, the PCA, in comparison to other dimension reduction methods, is more time-intensive and requires large memory consumption. Another drawback is that data standardization is recommended before using PCA. Performing PCA on unstandardized data would yield different results. PCA cannot be applied to nonlinear data due to its linear projections.

When PCA is used for tandem clustering, together with K-means, it can underperform. Indeed, Yeung and Ruzzo (2001) investigate that for gene expression data using PCA before clustering is not recommended. It is due to the components choice, that is made based on variability but still might miss some data structure information. However, still, PCA dimension reduction before segmentation is often used in practice, for instance in brain studies, where the data is high-dimensional (Kaya et al., 2017). Additionally, Soete and Carroll (1994) claim that by looking only at the first few components, the important structure of the data might be missed. They suggest using an

alternative method that is not sequential, but simultaneous. Reduced K-means is an approach that does dimension reduction and clustering at the same time. It recovers the true cluster design and performs well, also with many noise features.

Dimension reduction can also be done by ICS (Nordhausen and Ruiz-Gazen, 2022). ICS is a modern method that resembles PCA, but instead of one, it is based on the joint diagonalization of two scatter matrices, and eigenvalues are interpreted with regard to general kurtosis and not variance. It is arguable why a new subspace of reduced dimension should be created based on the large variance, like for PCA. Moreover, concerning the projection pursuit indicators, kurtosis is the most popular (Nordhausen and Ruiz-Gazen, 2022). It is that the projections of interest are those that do not resemble a normal distribution. Kurtosis then is useful for the identification of non-Gaussianity and hence, also for finding groups.

Research shows that the direction of the small kurtosis indicates large clusters and vice versa. Thus, invariant coordinates can be used for clustering (Fischer et al., 2017). According to Peña et al. (2010), the idea of minimizing the kurtosis to find clusters is more useful than maximizing the kurtosis direction, in case of two clusters of the same size. If the observations have the same distance to the mean, the variance would be almost zero, and the kurtosis would be small. Therefore, directions minimizing the general kurtosis will find clusters. For instance, Alashwali and Kent (2016) prove that in the context of the mixture of two groups, the clustering direction is the one spanned by the eigenvectors that belong to the smallest eigenvalue. Consequently, minimizing the general kurtosis finds big clusters, which is attractive for the cluster analysis.

3 Methodology

In this thesis, the number of clusters is assumed to be known and equal to K. We first perform dimension reduction. Dimension reduction is done by two different techniques simultaneously, namely PCA and ICS. Next, clustering is performed on the obtained subsets. We also use and compare two methods for cluster analysis: K-means and the EM algorithm. Parallelly, we perform clustering directly on the original dataset in order to compare findings.

The notation used in this paper is given below.

Notations

X	A p-multivariate real random vector.
X_n	A p-multivariate data with n observations.
x_i	A p-multivariate vector associated with the i^{th} observation.
p	Number of dimensions.
n	Number of observations.

3.1 Dimension Reduction

Two dimension reduction methods, namely PCA and ICS, are performed to get new subspaces that can be further used for clustering.

3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a non-parametric, unsupervised method that compresses data into a smaller set of variables that capture the essence of the original data. It is used to reduce dimensions and get uncorrelated variables (Hasan and Abdulazeez, 2021). PCA outputs Principal Components (PCs), new variables that are calculated based on the maximum variance. PCA computes eigenvalues for each component. The number of components equals the number of features. The PCA algorithm works in the following way:

1. Compute a p x p sample covariance matrix C.

2. Perform an eigenvalue decomposition $C = UDU^T$, where D is a diagonal p x p matrix with $d_{ii} > d_{jj}$, when $i > j$ and eigenvalues on the diagonal; and where columns of U are the eigenvectors (U is a p x p matrix).

3. Compute $Y = U^T X$, where Y's are the PCs, and X's are the features.

PCs are listed decreasingly, namely, the first component explains the most variation and the last one has the smallest variance. Therefore, PCA indicates that the first new variables are the best representation of the original data (Hasan and Abdulazeez, 2021). PCs are often chosen based on the percentage of cumulative variance they explain or based on the scree plots, where one is looking for the elbow. The elbow is a point at which the variance explained by each consecutive component reduces significantly. For the regular PCA with standardized data, the PCs with eigenvalues above

one are chosen. Moreover, since directions are sensitive to the scale of the data, in order to apply the PCA, data standardization has to be performed beforehand (Anowar et al., 2021).

3.1.2 Invariant Coordinate Selection

Invariant Coordinate Selection (ICS) is often called a generalized PCA (Caussinus and Ruiz-Gazen, 1995). Two methods are similar, however, instead of diagonalizing the variance-covariance matrix, ICS is based on the joint diagonalization of two scatter matrices (Nordhausen and Ruiz-Gazen, 2022).

Scatter Matrices Scatter matrices are the generalization of the covariance matrix. $S(X_n)$ is called a scatter matrix if it is affine equivariant. So if it follows:

$$S(X_n A + 1_n b') = A' S(X_n) A, \tag{1}$$

for $X_n = (x_1, ..., x_n)'$ p-variate data, where A is a full rank p x p matrix, 1_n a n-vector of ones and b a p-vector. It is true for any p x p $S(X_n)$ matrix that is symmetric and positive-definite. There is no universal rule on which scatter matrix to use, but it is known that their choice can impact the outcome. There are different classes of scatter matrices depending on how robust they are. Two ways of measuring robustness are the breakdown point and the influence function.

In robust statistics, a breakdown point is a point after which an estimator becomes ineffective. It is a proportion of outliers that the estimator can tolerate. It measures robustness, thus the higher the breakdown point, the better the estimator. For example, the breakdown point of a sample mean is $\frac{1}{n}$, with an asymptotic breakdown point of 0%. It means that only one outlier can break the estimator. Meanwhile, the asymptotic breakdown point of a sample median is the highest possible and equal to 50%. Another measure of robustness is the influence function. It resembles a derivative equation, but here the derivative is taken of the functional form of the estimator. It determines the change in the estimator when a small amount of contamination is present. For robust estimators, the influence function should be bounded (e.g., the influence function of sample mean is unbounded, but the influence function of sample median is bounded). The three classes of scatter matrices proposed by Tyler et al. (2009) are:

- Class I: not robust, the breakdown point is zero and the influence function is unbounded; sample covariance matrix belongs to this class.

- Class II: more robust than Class I, bounded influence function and positive breakdown point (but not larger than $\frac{1}{p+1}$); M-estimators belong to this class.
- Class III: the most robust class, high breakdown point; S-estimators and Minimum Covariance Determinant (MCD) estimator belong to this class.

In the presence of outliers, Class II or III scatter pairs would be advised (Tyler et al., 2009). In most cases, it is recommended to choose one scatter matrix from Class II and the second from Class III. However, the right choice of scatter matrices continues to be an open question in the literature. It is advised to try to implement different pairs of scatter matrices and then evaluate obtained ICS components (Tyler et al., 2009). Especially for outlier identification purposes, the first matrix has to be more robust. Then, the first components are maximizing the general kurtosis to find outliers; thus, investigating just the first components is enough. We will consider two cases: the first one with scatter matrices being COV & COV_4, which is a regular covariance matrix and a scatter matrix of fourth moments. COV_4 follows:

$$COV_4(X_n) = \frac{1}{(p+2)n} \sum_{i=1}^{n} d_i^2 (x_i - \bar{x})(x_i - \bar{x})', \tag{2}$$

where $\bar{x}$ is the empirical mean and $d_i^2 = (x_i - \bar{x})' COV(X_n)^{-1} (x_i - \bar{x})$ is the squared Mahalanobis distance. For the second case, we take more robust scatter matrices: MCD & COV. Minimum Covariance Determinant (MCD) is a highly robust and affine equivariant estimator of location and scatter. It uses a subset of data that obtains the smallest determinant of the covariance matrix. The scatter estimate is the covariance matrix taken on a subset, multiplied by a Fisher consistency correction (Hubert et al., 2018).

We choose those scatter pairs so that the first matrix per pair is more robust. Even though in the first case, the selected scatter matrices are from Class I, so not robust, there are many theoretical properties described in the literature that can be insightful (Tyler et al. (2009), Archimbaud et al. (2018b)).

ICS Fundamentals ICS aims to find a p x p transformation matrix $W(X_n)$ such that:

$$W(X_n) S_1(X_n) W(X_n)' = I_p \tag{3}$$

and

$$W(X_n)S_2(X_n)W(X_n)' = B(X_n). \tag{4}$$

$S_1(X_n)$ and $S_2(X_n)$ are scatter matrices and $B(X_n)$ is a diagonal matrix with eigenvalues in decreasing order on the diagonal. The rows of $W(X_n)$ are the equivalent eigenvectors. When all elements of $B(X_n)$ are different, the following equivariance property holds:

$$W(X_n)A^{-1} = JW(X_nA' + 1_nb'), \tag{5}$$

where J a p x p diagonal matrix that has ± 1 on its diagonal. By using a location functional T, we get:

$$W(X_n)(x_i - T(X_n)) = JW(X_nA' + 1_nb')((Ax_x + b) - T(X_nA' + 1_nb'). \tag{6}$$

Thus, $z_i = W(X_n)(x_i - T(X_n))$ are the ICS components that are affine invariant. The affine equivariance property of ICS means that the data can be rotated, scaled and such transformation will not affect the outcome. Hence, ICS components present the structure of the original data regardless of its initial coordinate system. This advantageous property makes ICS superior to PCA, which does not have such ability. PCs are invariant under orthogonal transformations but not under other affine transformations.

ICS for clustering In the context of clustering, we examine ICS property that enables finding groups in the data. First, the process of obtaining a lower-dimension subspace after transformation is explained. Once having selected the components, ICS can be used for cluster analysis.

After choosing the scatters, the ICS transformation is performed. The obtained diagonal elements of $B(X_n)$ matrix are the generalized kurtosis values of the ICS components. According to Tyler et al. (2009), the first and last eigenvalues of the transformation represent the general measurement of maximal/minimal kurtosis. We assume that the first and the last components that are not normally distributed, are of interest. To choose the components, we generalize the approach proposed by Archimbaud et al. (2018b), which considers normality tests. For each chosen pair of scatter matrices, we carry out the ICS transformation and then perform the D'Agostino test (DA) (Archimbaud et al., 2018b). Further, with the obtained p-values of the test, we start looking at the first and the last one, then comparing the second and one before the last, et cetera. It is based on the above-mentioned fact that the first and the last eigenvalues correspond to maximal and minimal generalized kurtosis, respectively. It differs from the component selection for outlier

identification purposes. When the outlier contamination is small, the components of interest are only the first ones (Archimbaud et al., 2018b). Maximizing the general kurtosis allows to find small clusters. Minimizing the general kurtosis indicates big clusters and the last components can display the structure of the data well. When choosing the components, we stop when the p-values are insignificant or when we know the number of true groups in the data. In that case, we stop with $K-1$ components. Thus, when investigating the components, we do not only start from the first ones, but at the same time, we check and compare the last components, and we go inwards.

The resulting lower-dimensional subspace resembles Fisher's linear discriminant subspace in the context of mixture of elliptical distributions, and thus can be used for clustering. Recalling the *Theorem* 3 (Tyler et al., 2009), when the data comes from the mixture of two multivariate normal distributions, with unequal means ($\mu_1 \neq \mu_2$) and with two proportional covariance matrices, that is Γ and $\lambda\Gamma$ (where $\lambda > 0$ and Γ belongs to symmetric positive matrices) we have three different cases after performing ICS transformation:

1. The first eigenvalue is larger than the others, and the other eigenvalues are equal to each other,
2. The first $p-1$ eigenvalues are equal and larger than the last eigenvalue,
3. All eigenvalues are equal.

When $p > 2$ and the first case is true, then the first eigenvector is proportional to $\Gamma^{-1}(\mu_1 - \mu_2)$. When the second case is true, the last eigenvector is proportional to $\Gamma^{-1}(\mu_1 - \mu_2)$. If the first eigenvalue is larger than the second for $p = 2$, then either the first or the last eigenvector is proportional to $\Gamma^{-1}(\mu_1 - \mu_2)$.

In all the above cases, mentioned eigenvectors resemble the Fisher's linear discriminant function. It is true, even when the class membership is unknown and it is generalized to a mixture of many elliptically symmetric distributions (Tyler et al., 2009). It allows us to find groups in the data. Thus, it is valuable for cluster identification purposes.

3.2 Clustering

We focus on two clustering methods, the K-means and the EM algorithm.

3.2.1 K-means

K-means is a simple, easy and fast to implement method. The name was firstly introduced by MacQueen (1967), but in fact, the method was discovered by Steinhaus (1956), under a different name: bagging predictors. It is an example of hard partitional clustering, namely it aims to divide the data into K partitions. The partitions are made based on the similarity measure (Euclidean distance). K is chosen in advance and it indicates the number of clusters. There are numerous ways to choose the K, one of which uses the elbow plot. The elbow is a point at which the within cluster sum of square obtained by each consecutive K reduces significantly. The assumptions of K-means are such that clusters have similar sizes and are spherical.

The objective of the K-means is to minimize the summed squared distance within clusters. Let C be the set of clusters, $C = C_1, C_2, ..., C_K$, we then aim to minimize:

$$\arg\min_{C} \sum_{i=1}^{K} \sum_{x \in C_i,} ||x - \mu_i||^2, \tag{7}$$

for $(x_1, x_2, ..., x_n)$ observations and for μ_i being the cluster C_i's center. The algorithm selects points for which it calculates the distance to other observations, it assigns the point to the nearest cluster and calculates the center of each cluster. Next, it reclusters and the process is repeated. Thus, the assignment step and the update step (means recalculation) are repeated until the cluster assignments do not change anymore (Likas et al., 2003).

3.2.2 Expectation Maximization

Expectation-Maximization (EM) is another clustering algorithm, firstly given its name and presented by Dempster et al. (1977). Unlike in the K-means, instead of selecting clusters to maximize the differences in the means, the EM algorithm computes membership probabilities. The method is based on two steps. The estimation step (E-Step) and the maximization step (M-Step). The E-step estimates the expected value of the latent variable and M-Step optimizes the parameters using maximum likelihood. The algorithm iterates until convergence. The EM clustering uses Gaussian probability distributions and is an example of soft clustering (Dempster et al., 1977).

Let $X = (x_1, x_2, ..., x_n)$ and $Z = (z_1, z_2, ..., z_n)$, Z being the latent variable (the unobserved variable that indicates group membership). If we have three clusters, as it is in our case of GMM we have: $X_i|(Z_i = 1) \sim N(\mu_1, \Sigma_1)$, $X_i|(Z_i = 2) \sim N(\mu_2, \Sigma_2)$ and $X_i|(Z_i = 3) \sim N(\mu_3, \Sigma_3)$, where the membership probabilities are: $P(Z_i = 1) = \pi_1$, $P(Z_i = 2) = \pi_2$ and $P(Z_i = 3) = \pi_3 = 1 - \pi_1 - \pi_2$. The aim of the algorithm is to estimate unknown parameters $\theta = (\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3)$. The E-step follows:

$$\Theta(\theta|\theta^{(t)}) = E_{Z|X,\theta(t)}[logL(\theta; X, Z)]. \tag{8}$$

Next, the M-step is performed such that:

$$\Pi^{(t+1)} = \arg\max_{\theta} \Theta(\theta|\theta^{(t)}). \tag{9}$$

The $logL(\theta; X, Z)$ used in equation 8 is the complete likelihood function:

$$logL(\theta; X, Z) = \prod_{i=1}^{n} \prod_{j=1}^{3} [f(x_i; \mu_j, \Sigma_j)]^{I[z_i=j]}, \tag{10}$$

where f is the multivariate normal density function and I is the indicator function. The indicator function takes the value equal to one if the equation $z_i = j$ is satisfied for a particular i and j, or zero otherwise. The algorithm iterates until convergence and the best fit values of the Gaussian parameters are achieved. If we know the parameters of the particular Gaussian, we can get the probabilities of coming from this or another Gaussian. Based on that we know which point came from where and the membership probabilities are obtained.

3.3 Evaluation

Two evaluation methods are used for cluster validation. An external validation ARI and an internal validation, the average silhouette coefficient.

3.3.1 ARI

Adjusted Rand Index (ARI) is a validation index that measures the similarity between two clustering results (Hubert and Arabie, 1985). To understand how it is computed, we first make use of the confusion matrix (Table 1). A confusion matrix, also called an error matrix, is used as a performance measure and shows the summary of predicted results of a classification task. TP indicates a positive prediction which in fact was true, thus TP is the number of true positives.

FP means that the prediction was positive while the actual value was negative. Hence, FP is the number of false positives (also known as Type 1 Error). Consequently, TN is the number of true negatives and FN false negatives (Type 2 Error). In the context of classification, a *Positive* condition means 1 (the same cluster) and a *Negative* means 0 (different cluster).

		Predicted Values	
		Positive	Negative
Actual Values	Positive	TP	FN
	Negative	FP	TN

Table 1: Confusion matrix.

ARI is an adjusted version of the Rand Index (RI). RI can be shown as a percentage of correctly clustered observations:

$$RI = \frac{TP + TN}{TP + FP + TN + FN}. \tag{11}$$

RI can also be represented by a formula:

$$RI = \frac{a + b}{\binom{n}{2}}, \tag{12}$$

where a is the count of pairs of points that belong to the same cluster in two clustering approaches, b is the count of pairs that are not grouped together and the denominator is the total number of pairs. Thus, the numerator of equations 11 and 12 represent the count of pairs in agreement.

ARI is adjusted for the number and size of clusters. If they are the same, under the randomness of cluster labels, its expected value is equal to zero. ARI is then:

$$ARI = \frac{RI - expected(RI)}{max(RI) - expected(RI)}, \tag{13}$$

where $expected(RI)$ is an expectation value of the Rand Index (Sinnott et al., 2016). ARI can yield negative values. Already a value of zero can indicate that clusterings are not similar to each other. The highest value ARI can yield is 1. Thus the higher the index number, the more similar two clusterings are.

3.3.2 Silhouette Coefficient

The average silhouette coefficient is a method proposed by Rousseeuw (1987) and it indicates how well observations were clustered. It calculates the silhouette score (silhouette width) for each observation using:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{14}$$

where $a(i)$ indicates the average dissimilarity of i to other elements in its cluster and $b(i)$ is the average dissimilarity to other elements in other clusters. The dissimilarity is the distance between two observations which indicates how different observations are from each other. The final score of the silhouette coefficient is then the average of all the scores. The coefficient can take values between -1 and 1, including the end values of the interval. A high value indicates that the observation was correctly grouped; thus a value close to -1 indicates misclassification. The scores are plotted on the silhouette graphs. The best clustering is when the silhouette heights are similar. As an example, Figure 1 shows two silhouette plots (Raymaekers and Rousseeuw, 2022). The left plot presents grouping into three clusters. The grey cluster's height is large and the width varies much. In this particular cluster, we see large scores (width) which indicate that points are located close to the center of a cluster, but also we see very low scores. This means that some points in the grey cluster are close to other clusters. Next, we look at the right graph with four partitions. Here, the heights are almost the same, and silhouettes look similar. Even though negative values (wrongly classified observations) are visible, the overall average silhouette width is larger than for the left graph.

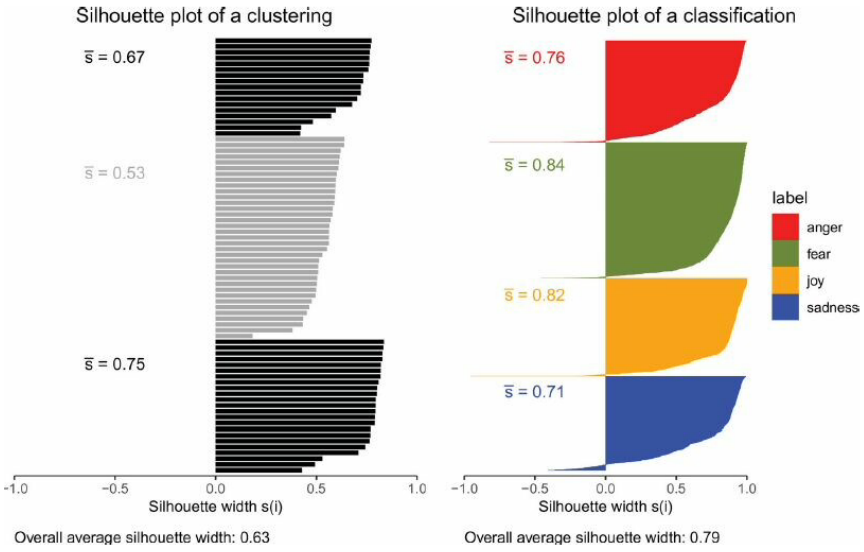

Figure 1: Example of silhouette plots.

4 Simulations

In this section, the simulation design and results are described.

4.1 Data Design

We are performing simulations to account for the different grouping design. We simulate two Gaussian Mixture Models (GMM) based on different observations' split:

- **Model 1**: data split - 50%, 30% and 20%
- **Model 2**: data split - 50%, 40% and 10%

As there is no universal rule of thumb on minimum sample size for clustering purposes (Siddiqui, 2013), we choose to have 1000 observations. The data follows:

$$\begin{aligned} X &\sim (1-\epsilon)N(0,\Sigma) + \epsilon_1 N(\delta e_1,\Sigma) + \epsilon_2 N(\delta e_2,\Sigma) \\ &\sim (1-\epsilon)N(\mu_0,\Sigma) + \epsilon_1 N(\mu_1,\Sigma) + \epsilon_2 N(\mu_2,\Sigma) \end{aligned} \tag{15}$$

where δ is equal to 8, $\epsilon = \epsilon_1 + \epsilon_2$, and $X = (X_1, ..., X_p)'$ is a p-multivariate random vector and the distribution of X is a mixture of three Gaussian Distributions. We have different location measures; μ_0 is a vector of zeros, μ_1 is a first unit vector multiplied by δ, and μ_2 is a second unit vector multiplied by δ. For the above model, we consider two scenarios for positive-definite covariance matrix Σ: the identity matrix and the diagonal matrix ($\Sigma = diag(1, 1, \delta, ..., \delta)$).

Table 2 shows the simulation set-up summary. Each case is performed for three different dimensions. There are three scenarios for the number of dimensions: 10, 50 or 100 ($p_1 = 10, p_2 = 50, p_3 = 100$, respectively). In this set-up, three groups are visible in the data. To hide it for further analysis, we rotate the data with an orthogonal matrix (Fischer et al., 2017). We use the Toeplitz matrix for this, which has each descending diagonal element constant. For the dataset, we are also adding a categorical variable *group* that indicates from which group the observation is.

	Model	Data Split	Sigma
Case 1	1	50%, 30%, 20%	Diagonal
Case 2	1	50%, 30%, 20%	Identity
Case 3	2	50%, 40%, 10%	Diagonal
Case 4	2	50%, 40%, 10%	Identity

Table 2: Data simulation set-up.

Simulation is performed in R and required packages are mentioned below. 100 samples are generated with the number of observations equal to 1000 and the dimension p of 10, 50 and 100. Data follows GMM with various observations' split and different covariance matrices (see Table 2 for details). Before we present the results, we explain in detail how the simulation was performed for original and modified datasets.

Simulation design per dataset For Original Data, for each of the four cases and three different dimensions, we create simulation functions. Clustering is done directly on the data generated by GMM. For the K-means clustering, we use the *cluster* package (Maechler et al., 2019) and we take the true number of groups, namely three. To perform EM clustering we use the *mclust* package (Scrucca et al., 2016), also with the group number equal to three. For each clustering algorithm, we calculate the ARI and we take an average over all simulations. ARI tells us how well the data was clustered comparing the classification output to the true group membership.

For PCA Modified Data, we apply PCA for dimension reduction before performing cluster analysis. We make use of the *factoextra* package (Kassambara and Mundt, 2020). PCA output shows eigenvalues and we take those into account when deciding how many components to take. We take those components that have corresponding eigenvalues larger than one (Jinga et al., 2006). Because we know the number of groups is three ($K = 3$), it follows that the maximum number of chosen principal components is equal to two (max $K - 1$ components) (Ding and He, 2004). With the new subspace, we proceed with clustering using the same approach as for the Original Data. Additional to the ARI results, we also report the average number of selected components.

For ICS Modified Data, the dimension reduction is done by ICS. We use the *ICSOutlier* package (Archimbaud et al., 2018c), in which we can specify scatter matrices. We have two different functions per each case of simulation. This is because we first perform ICS with scatter matrices being COV & COV_4, and then with more robust scatter pair MCD & COV. To choose the

invariant coordinates we use the normality test, namely the D'Agostino test (Archimbaud et al., 2018b), and we set the significance level to 5%. The idea is such that the new subspace consists of not normal components. We start by considering the p-values of the first and last component, next we look at the second and one before the last. We check if they are smaller than the significance level, and we choose the one that is the lowest. As we have three groups in the data, the maximum number of coordinates is two. Figure 2 plots the first and the last component after performing ICS transformation with MCD & COV as scatters. Three groups are clearly visible on the plot. For all new subspaces created by ICS, we want to find two components with clearly visible groups.

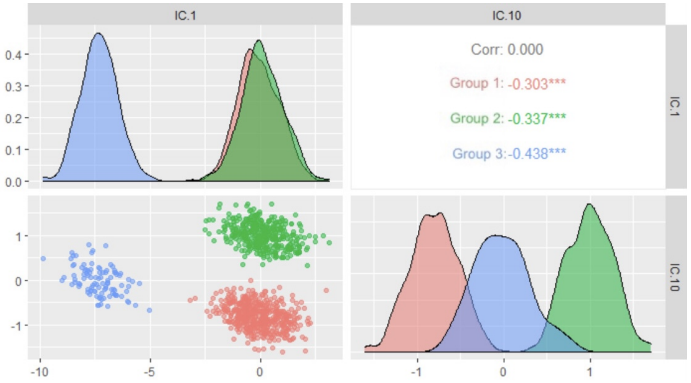

Figure 2: First and last component for $Case3$ when using MCD & COV.

When performing ICS with MCD & COV, the default for the MCD's *alpha* is 50%. This is the breakdown point and this parameter is controlling the size of the subset that is used for the determinant minimization. We use this set-up as a foundation for our ICS MCD & COV analysis. However, we encounter problems with coordinates selection and change the parameters for certain cases. The problems with MCD will be mentioned later. The *alpha* can take values from 0.5 to 1, so we also apply MCD *alpha* 75% to see if it performs better (namely, if it chooses, on average, two components). Another modification to the above-mentioned MCD is the *nsamp*, the number of subsets that is used for primary estimates. The default is equal to 500. However, we also check the *deterministic* MCD, where the h central observations of six deterministic estimators are the start (Hubert et al., 2012). The combination of *deterministic* and *alpha* 0.75 is also applied to examine if it recovers the right number for components. Lastly, similarly as for the PCA approach, we perform clustering on the new subspace and obtain mean ARI results.

As a reference to our analysis, we review the *ICSShiny* package (Archimbaud et al., 2018a) with the ICS application. The application visualizes coordinates for different scatter matrices and investigates more normality tests for the coordinates selection.

Plots description Below graphs show the results of simulations of the four cases. The red color indicates the average ARI value, the color green indicates the minimum ARI, and the blue is the maximum ARI. Color coding is the same for all graphs. In the x axis, we can see five different methods, which are the methods the data was modified with. PCA is the data modified with PCA algorithm, Original Data is unmodified. ICS1 is the ICS with COV & COV_4, ICS2 is the ICS with MCD & COV, ICS3 is the ICS with modified MCD & COV. Adjustments to MCD in ICS3 are described in detail per case.

4.2 Case 1

For $Case1$, the ICS3 uses $\alpha = 0.75$ for MCD. When the dimension is equal to 10, $\Sigma = diag(1, 1, 8, ..., 8)$ and K-means is the clustering method, we see that the highest mean values were received by using ICS2 and ICS3 (Figure 3). Although their minimum and maximum values have big differences, the medians are also the highest (See Table 6 in Appendix B). For the EM clustering results, we see that ICS2, ICS3 and Original Data have similar results, while PCA performs significantly worse than other methods (Figure 4). It is in line with PCA fundamentals, that it looks at the overall largest data variability but cannot account well for variability that is put only on the last variables. The median value for ARI when EM is used is the same for ICS2, ICS3, and for Original Data (Table 6 in Appendix B). Here (when $p = 10$), for K-means and EM clustering, the preferred dimension reduction method is ICS3 as it is fast, has the best results for ARI, and always finds two components (Table 6 in Appendix B).

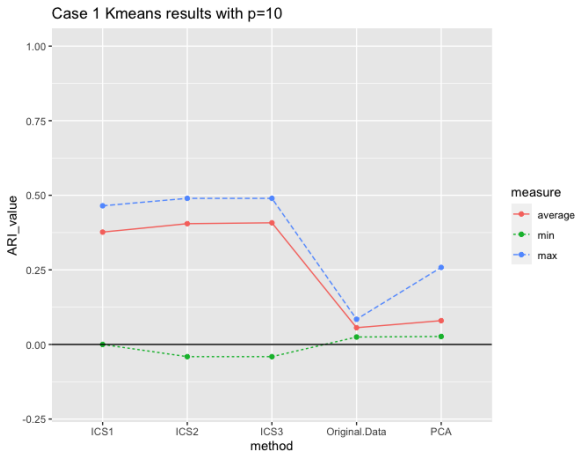

Figure 3: Simulation K-means results for Case 1, dim=10.

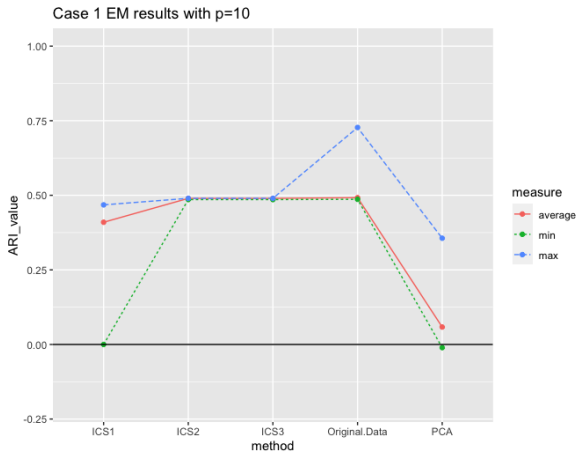

Figure 4: Simulation EM results for Case 1, dim=10.

For higher dimensions, $p = 50, 100$, we see from the graphs that all ARI scores are small, around zero, even negative (Figures 5,6,7,8). This indicates misclassification. The explanation for ICS1 not performing well is the fact that the observation split for $Case1$ consists of one group that is around 20% of all data. It has been proven that it is a limiting case for COV & COV_4 scatters (Archimbaud et al., 2018b). There are no such proofs for MCD & COV due to advanced and extensive mathematical calculations, but it might also be the reason why ICS2 and ICS3 are not performing well for this case. For higher dimensions of $Case1$, we do not have satisfactory results. Moreover, an observation about the performance of PCA is such that it obtains slightly higher average ARI results, and it always chooses two components for its clustering subspace (See Table 6 in Appendix B).

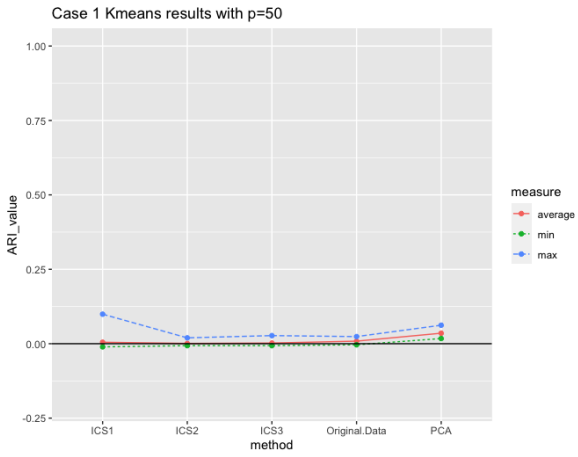

Figure 5: Simulation K-means results for Case 1, dim=50.

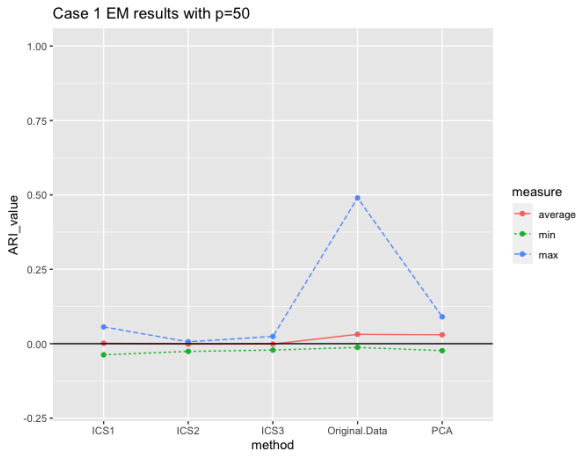

Figure 6: Simulation EM results for Case 1, dim=50.

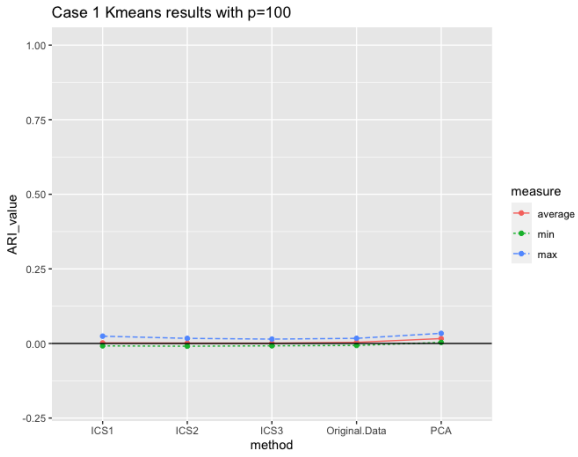

Figure 7: Simulation K-means results for Case 1, dim=100.

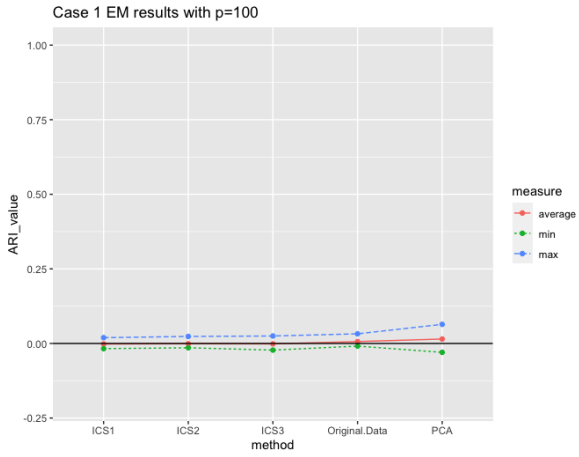

Figure 8: Simulation EM results for Case 1, dim=100.

4.3 Case 2

For $Case2$, the ICS3 uses $\alpha = 0.75$ for MCD.

$Case2$ indicates that the covariance matrix is the identity matrix, and the observation split is 50%,30%, and 20%. When K-means is used as a clustering method for 10 dimensions, it is noticeable that Original Data has the smallest mean ARI value. Other methods obtain better results (Figure 9). ICS2 and ICS3 find two coordinates (Table 7 in Appendix B) and have the same median values. PCA, however, does not find two coordinates as the mean for the number of components is 1.96 (Table 7 Appendix B). When EM clustering is used, we see an opposite trend to the K-means. Figure 10 shows that Original Data obtained the highest mean ARI value, whereas results for ICS methods are much smaller. Further, PCA's mean value is high but it does not always find two components (Table 7 Appendix B). ICS methods have the biggest value amplitude (around 0.5) for K-means, while for EM it is almost zero for ICS2 and ICS3. Even though ICS3 does not reach Original Data's mean ARI value, it chooses the two components well, thus we recommend using ICS3 as a base for both clustering methods when there are ten dimensions.

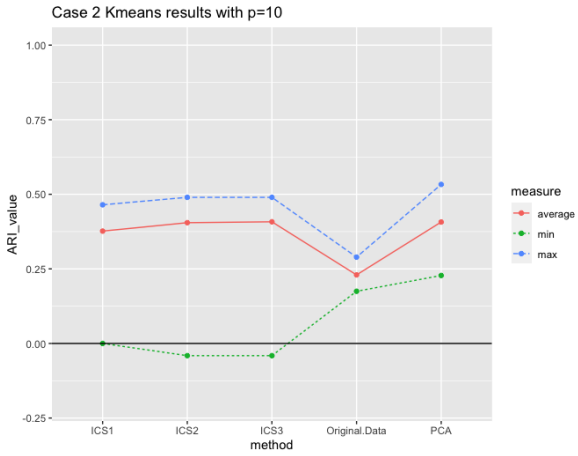

Figure 9: Simulation K-means results for Case 2, dim=10.

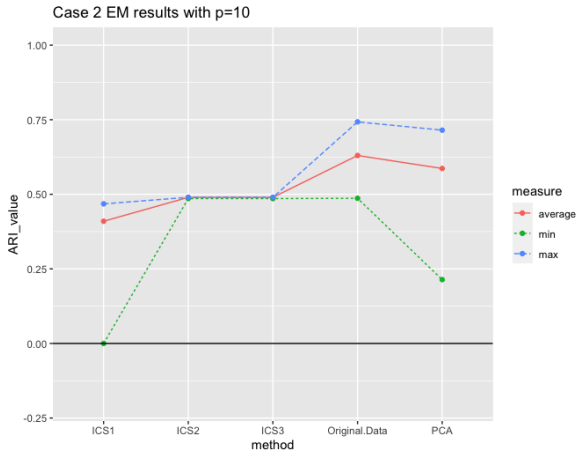

Figure 10: Simulation EM results for Case 2, dim=10.

Figure 11 for results of dimensions 50 and Figure 13 for results of dimensions 100 look alike. In both cases, K-means is used for clustering and the best results are obtained for PCA. As mentioned above, ICS is not performing well in case of a data split with a group of 20%. Similarly, for EM clustering results (Figures 12 and 14), ICS methods do not perform well and they do not choose two components (Table 7 Appendix B). However PCA, with this simulated data and the covariance matrix being identity, it always finds two components. Therefore, for dimensions equal to 50 and

100, the PCA obtained subspace should be used for clustering.

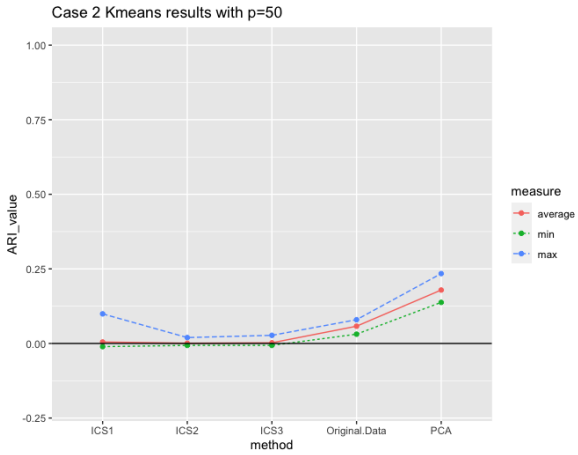

Figure 11: Simulation K-means results for Case 2, dim=50.

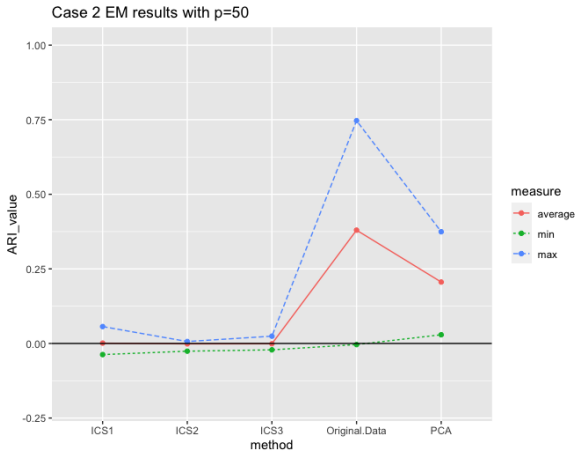

Figure 12: Simulation EM results for Case 2, dim=50.

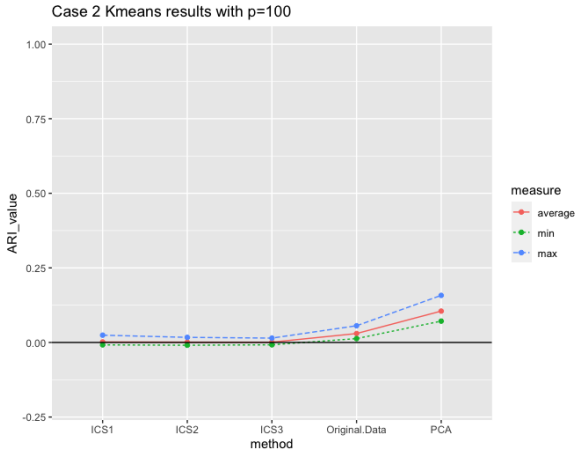

Figure 13: Simulation K-means results for Case 2, dim=100.

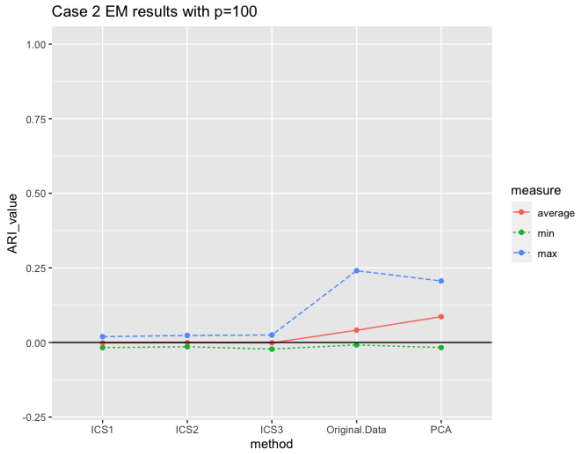

Figure 14: Simulation EM results for Case 2, dim=100.

4.4 Case 3

In $Case3$, for dimensions $p = 10, 50$ the ICS3 uses $\alpha = 0.75$ for MCD. On average, we see high ARI numbers (larger than 0.5) for ICS for both K-means and EM (Figures 15 and 16). $Case3$ has observations split of 50%,40%, and 10% thus, the ICS transformation with COV & COV_4 scatters should not encounter any problems. Moreover, all ICS methods choose two coordinates (Table 8 Appendix B). As the mean ARI value is the largest for ICS1, we conclude that it is an optimal choice for $Case3$ clustering when $p = 10$. The EM clustering results also obtain high values for ICS methods and for the Original Data. The interval of min/max ARI values for Original Data is

significantly smaller than for other methods, thus the EM clustering should be done on unmodified data when $p = 10$. The low performance of PCA is noticeable for both EM and K-means results. It is due to covariance matrix being $\Sigma = diag(1, 1, 8, ..., 8)$.

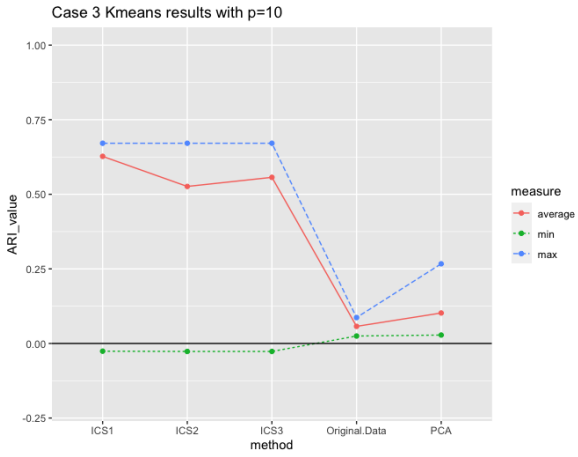

Figure 15: Simulation K-means results for Case 3, dim=10.

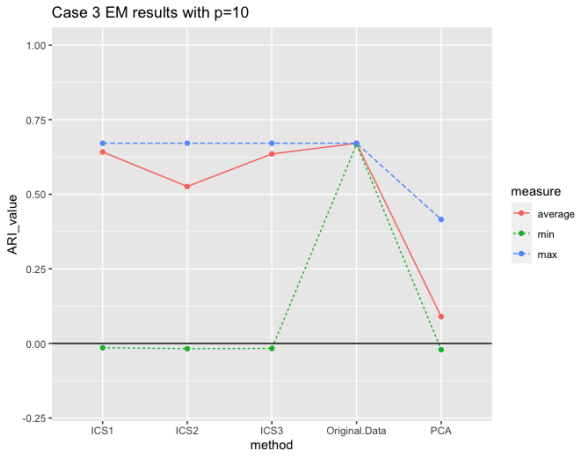

Figure 16: Simulation EM results for Case 3, dim=10.

Rather small numbers are obtained for average ARI results of K-means and EM clustering for $p = 50$ (Figures 17 and 18). PCA is finding two components, but none of ICS methods obtain two on average (Table 8 Appendix B). Therefore, in this case, clustering should be performed directly on the Original Data, as none of the dimension reduction methods significantly improve the clustering performance.

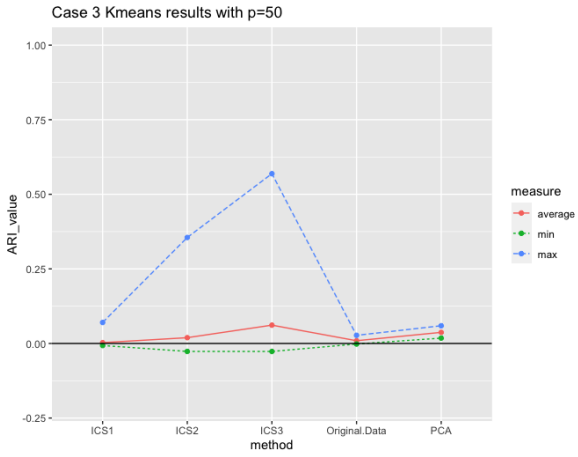

Figure 17: Simulation K-means results for Case 3, dim=50.

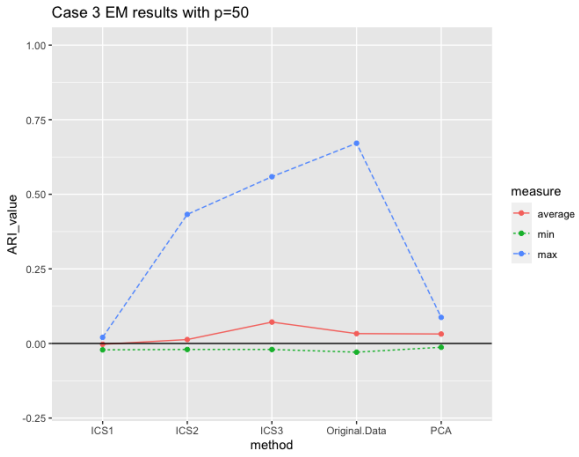

Figure 18: Simulation EM results for Case 3, dim=50.

Below graphs (Figures 19 and 20) show the results of $Case3$ with $p = 100$. ICS3 uses here a

deterministic version of MCD. Again, the values obtained by ICS are, on average low and do not find two components (Table 8 Appendix B). PCA subspace gives higher mean ARI with always two principal components as its base.

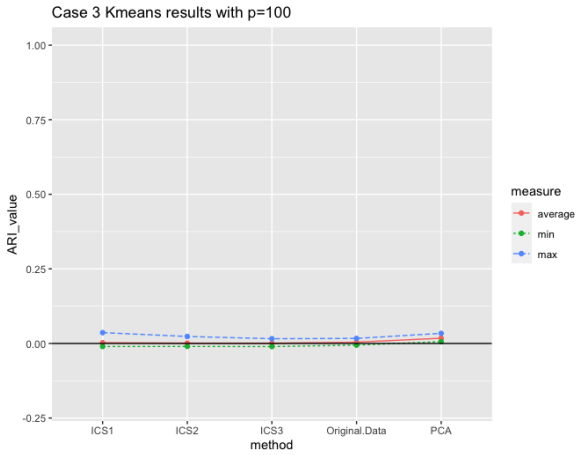

Figure 19: Simulation K-means results for Case 3, dim=100.

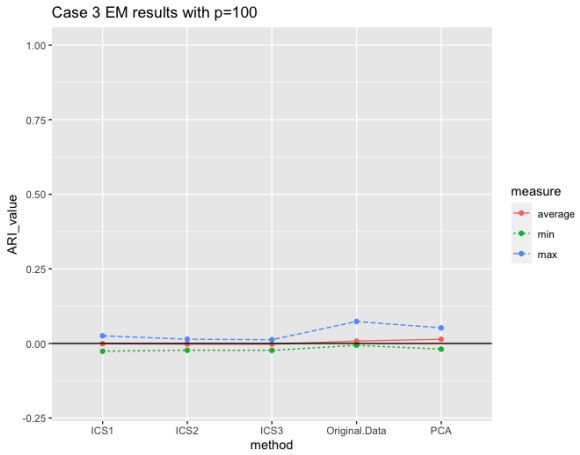

Figure 20: Simulation EM results for Case 3, dim=100.

4.5 Case 4

When $p = 10$ in $Case4$, ICS3 uses $\alpha = 0.75$ for MCD. Figure 21 shows the outcomes of the K-means clustering. Original Data's results are substantially smaller than the rest. $Case4$ is a set-up where we have the identity covariance matrix and the data split of 50%,40%, and 10%. Therefore, PCA should not encounter any problems with choosing the right components, and ICS transformation should perform well. Table 9 in Appendix B shows that indeed in all the cases, the average of chosen components is equal to two. Looking at the obtained mean ARI values, we see that ICS1 has the highest number. Even though the min/max interval of ARI values is large, the median is equal to 0.6682, which is significantly larger than the Original Data and similar to other ICS outputs (See Table 9 in Appendix B). Figure 22 presents the results of EM clustering, which are on average high for all methods (around 0.64). PCA and ICS correctly choose two components. Nevertheless, the EM clustering on Original Data obtains the best ARI value and has the smallest min/max interval.

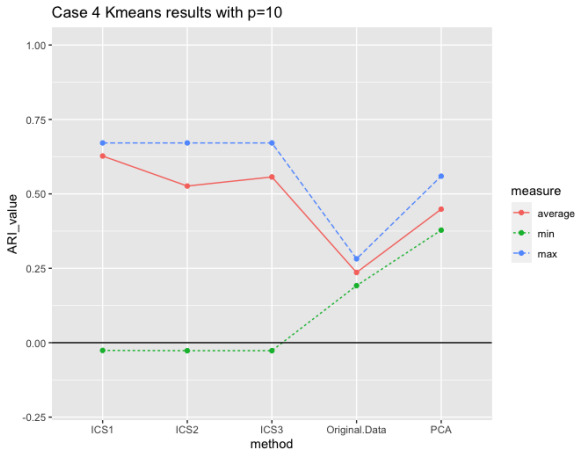

Figure 21: Simulation K-means results for Case 4, dim=10.

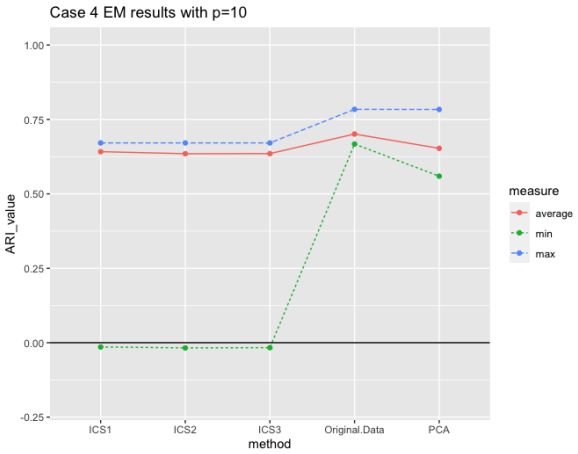

Figure 22: Simulation EM results for Case 4, dim=10.

For dimensions equal to 50 and 100 (for *Case*4) the ICS3 uses the *deterministic* version of *MCD*. Figures 23, 24, 25 and 26 present ARI results. Surprisingly, ICS methods do not work well, with the average ARI numbers around zero. Additionally, ICS methods choose at most 1.6 components, while PCA finds exactly two (See Table 9 in Appendix B). In almost all cases, PCA obtains the best mean results.

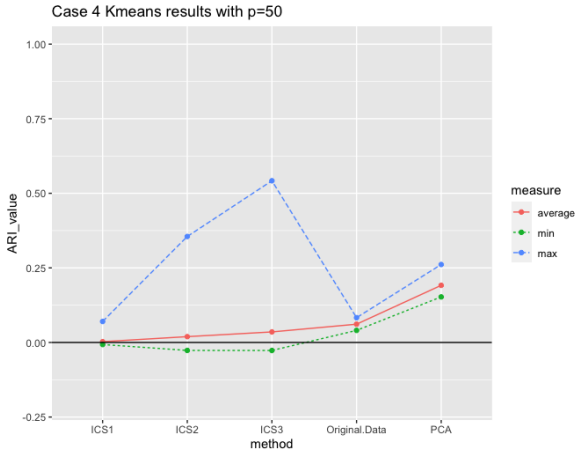

Figure 23: Simulation K-means results for Case 4, dim=50.

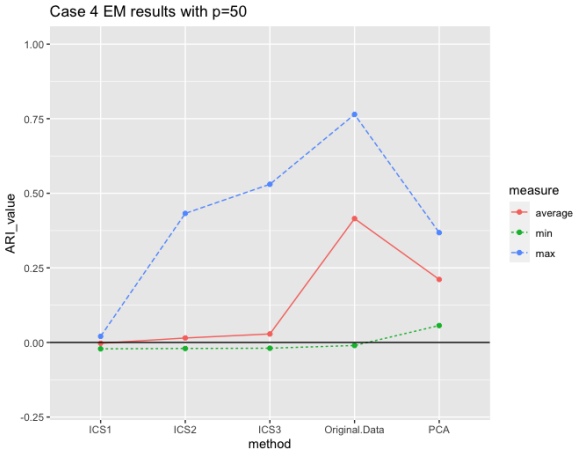

Figure 24: Simulation EM results for Case 4, dim=50.

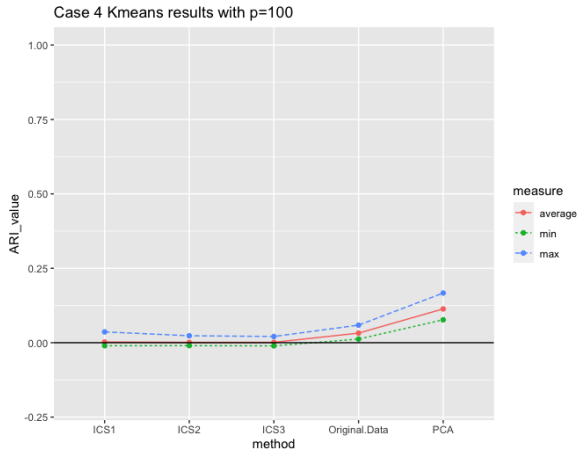

Figure 25: Simulation K-means results for Case 4, dim=100.

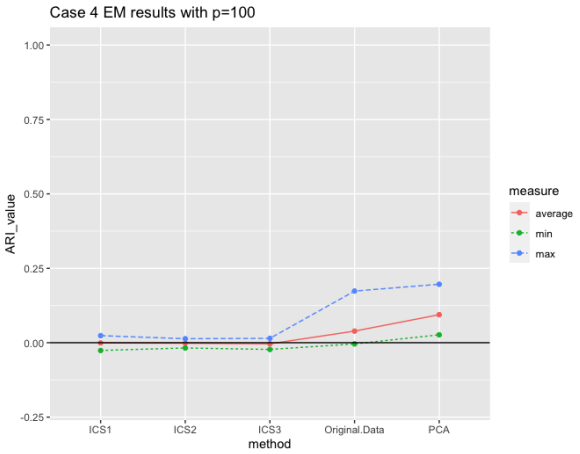

Figure 26: Simulation EM results for Case 4, dim=100.

5 Real Dataset

For completeness, after simulation analysis, we apply the same process to the real dataset. The *Vehicle* dataset is a built-in dataset in R package called *mlbench*, and the data comes directly from the UCI repository (Newman et al., 1998). The dataset consists of 846 observations and 19 variables. All variables are numerical but one, which is a categorical variable indicating the type of a vehicle.

The variables are the features that were extracted from the silhouettes by the Hierarchical Image Processing System (see Table 5 in Appendix A for features description). As we have a *Class* feature, we know exactly the true groups of vehicles. There are four types, and the exact membership can be seen in the table below (Table 3). Group identification was designed in a way that the three groups would be easily distinguishable (bus, van, car), making it more difficult to recover the right car type (*opel* or *saab*).

Group	Type	Class	Class %
Bus	Bus	218	26%
Car	Opel	212	25%
Car	Saab	217	26%
Van	Van	199	24%
	Total	846	100%

Table 3: Descriptives of *Class* variable.

We apply the same methodology to the *Vehicle* dataset. First, the clustering is done directly on the data (only the *Class* feature is deleted from the dataset, and it is treated as a true group vector, which is later compared to the clustering results). As a result, we get ARI values for K-means and EM algorithms. Both of them are similar, around 0.07 (Table 4).

	ARI Kmeans	ARI EM	Comp. Nr.
Original Data	0.0726	0.0796	-
PCA	0.0766	0.0477	3
ICS1	0.0207	0.0139	3
ICS2	0.0569	0.0732	3

Table 4: Real dataset results.

Next, we apply PCA to reduce the dimension. The first four principal components have a corresponding eigenvalue larger than one. Those four components would have been chosen as a new subset, however, in this case, we know the true number of car groups. Thus, we can select maximum of three components to be our base for clustering. The graph below shows how much variance is explained with a given number or components (Figure 27). Three principal components explain 80% of the total variance. For clustering, we know that the true number of clusters is four. In order to verify, we check the average silhouette coefficient for K-means and it also gives us the optimal number of clusters being equal to four (Figure 28). The obtained ARI value for K-means is 0.0766, and it is larger than the Original Data. However, the ARI result for EM clustering is equal to 0.0477, which is smaller than the original one (Table 4).

For ICS, to choose coordinates, we apply the same logic as in the simulated data. We do the normality test and check the obtained p-values. However, in this case, we choose the significance level to be 1%, as 5% would lead to too many components. For ICS1 (when COV & COV_4 are the scatters) we have that the first four coordinates are significant. Because we have four groups in the data, we can have at most three coordinates; thus, we take the first three to be our new subspace. In this case, the p-value of the last component is already larger than 0.01, thus insignificant. Figure 29 shows chosen components, the first three as mentioned above. It is noticeable that their densities are almost the same. Namely, it is hard to observe the vehicle's groups. Looking at IC.1 & IC.3, and IC.2 & IC.3, the *bus* class (in red) is more distinguishable from the others. Moreover, the ARI values obtained for ICS1 are much smaller than the rest. The scatter choice of COV & COV_4 is not performing well. When we look at the data division into four classes (Table 3), we see that

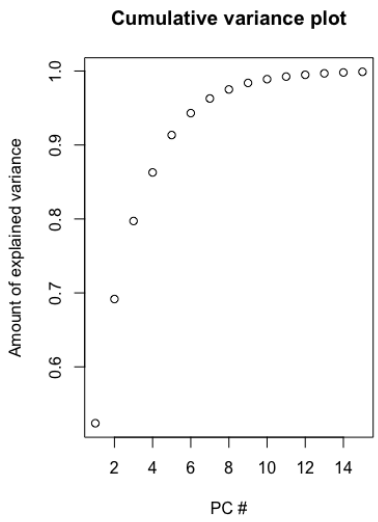

Figure 27: Cumulative variance plot for PCA on *Vehicle* dataset.

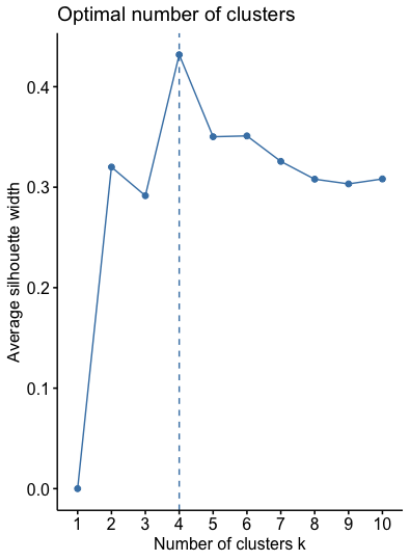

Figure 28: Optimal K for K-means performed on the PCA subset.

vehicle groups are around 25%. This might be the reason why ICS1 is not working properly.

When using MCD & COV as scatters (ICS2), we obtain six first components to be significant, but also the last one. Next, the p-values are compared and three coordinates are chosen (first, second, and the last). Figure 30 shows selected components. The density of IC.1 and IC.2 are almost the same, while the plot of IC.18 displays different densities. As mentioned previously, the data was designed such that three groups will be easily distinguishable (bus, van, car), making it more difficult to recover the right car type (*opel* or *saab*). Indeed, when looking at the IC.1 and IC.18 plot, we see the purple *van* observations on the left, almost all red *bus* observations on the right, and in the middle, we see green and blue ones. Thus, the groups are distinguishable, but not easily, as we see some overlaps. In this regard, ICS2 is performing better than ICS1. Clustering results with ICS2 are higher for EM and almost the same as for the Original Data. (Table 4).

All our approaches follow the tandem clustering form. As mentioned in the Literature (section 2), there exists an alternative method that is not done sequentially but simultaneously performs dimension reduction and clustering, namely the Reduced K-means. We apply it by using the *clustrd* package (Markos et al., 2019). We set the number of clusters to four and the dimensions of a solution, that is the new subspace, to three. With an obtained new clustering, we compute the ARI index. ARI equals 0.0761. It is similar to the values achieved with the Original Data (Table 4), but significantly larger than for all ICS cases. When comparing the tandem clustering of

PCA and K-means with the Reduced K-means, we obtain similar numbers. The Reduced K-means is only smaller by 0.0005. However, the Reduced K-means performs significantly better than the tandem clustering with the EM algorithm (Table 4).

In fact, the ARI results are not high, as the largest value is only 0.0796. Thus, if the K-means is the desired clustering method, we suggest performing the analysis on the PCA modified data, or using the Reduced K-means approach. However, when EM is used, we suggest modifying the data with ICS2 or by simply using the original $Vehicle$ dataset. Moreover, we notice that the distribution of the data is skewed. As ICS requires elliptical distribution, this might be the reason of clustering not performing well.

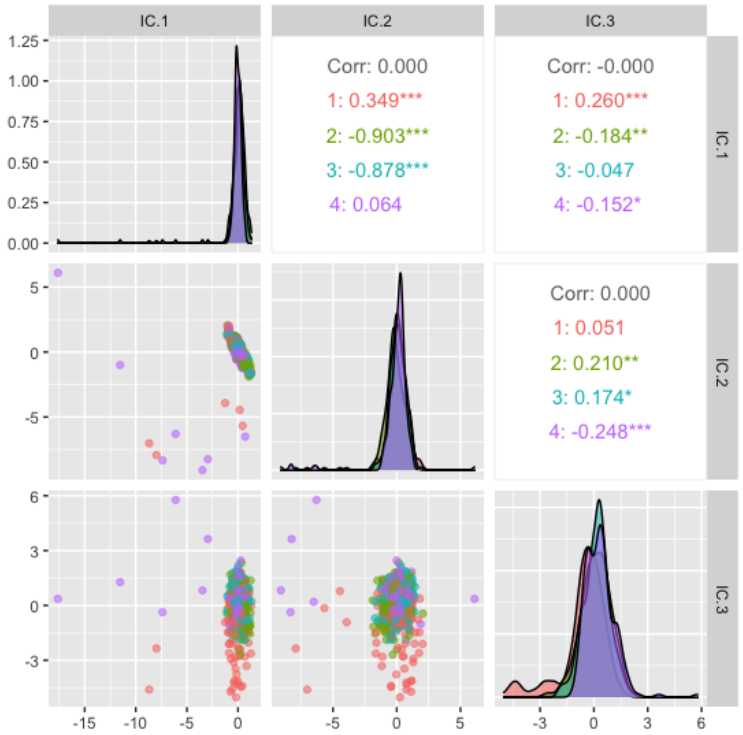

Figure 29: Invariant Coordinates chosen by ICS1. The class coding is: 1 = *bus*, 2 = *opel*, 3 = *saab* and 4 = *van*.

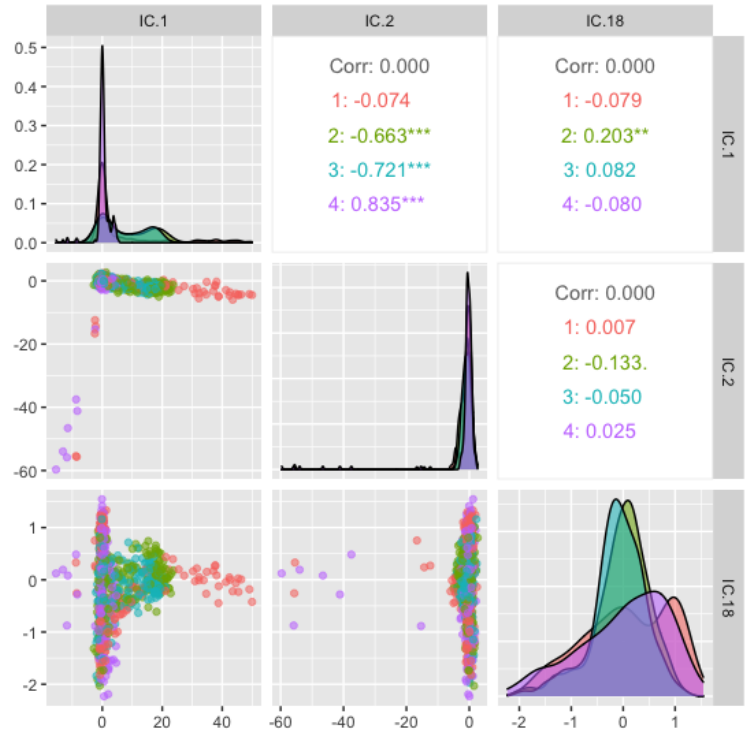

Figure 30: Invariant Coordinates chosen by ICS2. The class coding is: 1 = *bus*, 2 = *opel*, 3 = *saab* and 4 = *van*.

6 Conclusion and discussion

The main objective of this thesis is to improve clustering identification by first performing dimension reduction. Shrinking the feature space is beneficial for many complex algorithms as it increases accuracy and efficiency and simplifies data interpretation (Anowar et al., 2021). Various examples containing PCA and K-means algorithms have already been studied (Kaya et al. (2017), Lee et al. (2009)). However, PCA for tandem clustering might be inefficient. It is due to reducing the dimension only based on the variability and not finding the true structure of the data (Soete and Carroll, 1994). Therefore, we proceed with involving ICS in order to enhance clustering with dimension reduction in a sequential way. The ICS method focuses on finding the structure of the data and it considers the general kurtosis. Maximizing and minimizing the general kurtosis, allows to detect small and large clusters, respectively. ICS property of finding Fisher's linear discriminant subspace, in the context of mixture of elliptical distributions, enables one to use it for cluster analysis.

We perform clustering on simulated and real datasets. In both cases, we assume the number of clusters is known. Simulated data is generated by GMM with two set-ups for the covariance matrix and two models for the observations split. We perform K-means and EM clustering on the unmodified data but also on the data with dimensions reduced by PCA and ICS. Each analysis is applied for three different numbers of variables (dimensions equal to 10, 50 and 100) and diverse variants of ICS.

"Does the ICS dimension reduction before clustering improve cluster identification?" The answer to our research question is ambiguous and depends on the used data. Our observations and advice for further research are discussed below.

We have seen that, in general, when Model 2 is used ($Case3$ and $Case4$), the ICS modified data yields good results. When the dimension is equal to 10, ICS methods always choose the right number of components, and the clustering validation index obtains high values. When K-means is the clustering algorithm, ICS dimension reduction improves cluster analysis. For the EM, using both ICS and unmodified data obtain good results. However, for higher dimensions ($p = 50, 100$), we encounter problems while applying ICS. The average number of chosen components is not equal to two namely ICS is not always finding the right coordinates. To correct for it, we set the significance level to 5% and adjust MCD when it is used as a scatter for ICS. Nevertheless, it still does not produce optimal results. For $Case1$ and $Case2$, when the observations' split is

50%,30%,20% (Model 1), ICS with COV & COV_4 is not functioning well. Additionally, ICS with MCD & COV is also not performing well. Even when the adjustments are made for the MCD, ICS is not always choosing the right number of components. Consequently, the ICS is not improving the cluster analysis. For further research, we suggest applying various normality tests to choose the invariant coordinates. We only use the D'Agostino test but perhaps verifying other normality tests will provide the right components, which can yield better clustering results.

Although for the $Vehicle$ dataset we obtained the right number of components, the ICS ARI results do not exceed Original Data values. An additional investigation should also consider other scatter matrices for ICS. Possibly, those can have a positive effect on the ICS performance and thus improve clustering. To enhance components selection, the following analysis should also inspect adjusting p-values for multiple testing. Accounting for the distribution of the data shall also be considered. For instance, taking a *log* transformation of the data should solve the poor ICS performance. For the $Vehicle$ dataset we also apply the Reduced K-means. This is an alternative to tandem clustering, as dimension reduction and clustering are done at the same time. We obtain better results than those for ICS, thus we suggest performing the Reduced K-means on the simulated data as well.

Lastly, regarding time efficiency, the clustering done on the ICS modified data was much faster. For instance, the code for Original Data with 100 dimensions was running for 1.5 days, while for ICS only about 20 minutes ($Case3$). Additionally, due to time efficiency and many scenarios with a different number of dimensions or other data variations, our simulation was over 100 samples. We suggest that further study should generate at least 1000 samples for the GMM data.

References

Alashwali, F. and Kent, J. T. (2016). The use of a common location measure in the invariant coordinate selection and projection pursuit. *Journal of Multivariate Analysis*, 152:145–161.

Anowar, F., Sadaoui, S., and Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, 40.

Archimbaud, A., May, J., Nordhausen, K., and Ruiz-Gazen, A. (2018a). *ICSShiny: ICS via a Shiny Application*. R package version 0.5.

Archimbaud, A., Nordhausen, K., and Ruiz-Gazen, A. (2018b). ICS for multivariate outlier detection with application to quality control. *Computational Statistics Data Analysis*, 128:184–199.

Archimbaud, A., Nordhausen, K., and Ruiz-Gazen, A. (2018c). *ICSOutlier: Outlier Detection Using Invariant Coordinate Selection*. R package version 0.3-0.

Caussinus, H. and Ruiz-Gazen, A. (1995). Metrics for finding typical structures by means of principal component analysis. *Data science and its Applications*, pages 177–192.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on Machine learning*, page 29.

Ding, C., He, X., Zha, H., and Simon, H. D. (2002). Adaptive dimension reduction for clustering high dimensional data. *2002 IEEE International Conference on Data Mining, 2002. Proceedings*, pages 147–154.

Fischer, D., Honkatukia, M., Tuiskula-Haavisto, M., Nordhausen, K., Cavero, D., Preisinger, R., and Vilkki, J. (2017). Subgroup detection in genotype data using invariant coordinate selection. *BMC bioinformatics*, 18(1):1–9.

Hasan, B. M. S. and Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, 2(1):20–30.

Hennig, C. (2022). An empirical comparison and characterisation of nine popular clustering methods. *Advances in Data Analysis and Classification*, 16(1):201–229.

Hubert, L. and Arabie, P. (1985). Comparing clusterings. *Journal of Classification*, 2:193–218.

Hubert, M., Debruyne, M., and Rousseeuw, P. J. (2018). Minimum covariance determinant and extension. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(3):e1421.

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3):618–637.

Jinga, X. Y., Wong, H. S., and Zhang, D. (2006). Face recognition based on 2D Fisherface approach. *Pattern Recognition*, 39(4):707–710.

Kassambara, A. and Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

Kaya, I. E., Pehlivanlı, A. C., Sekizkardes, E. G., and Ibrikci, T. (2017). PCA based clustering for brain tumor segmentation of t1w mri images. *Computer Methods and Programs in Biomedicine*, 140:19–28.

Kinnunen, T., Sidoroff, I., Tuononen, M., and Fränti, P. (2011). Comparison of clustering methods: A case study of text-independent speaker modeling. *Pattern Recognition Letters*, 32(13):1604–1617.

Lee, C., Abdool, A., and Huang, C.-H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics*, 10(1):1–13.

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

Ma, J. and Yuan, Y. (2019). Dimension reduction of image deep feature using PCA. *Journal of Visual Communication and Image Representation*, 63.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *In 5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2019). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.0 — For new features, see the 'Changelog' file (in the package source).

Markos, A., Iodice D'Enza, A., and van de Velden, M. (2019). Beyond tandem analysis: Joint dimension reduction and clustering in R. *Journal of Statistical Software*, 91(10):1–24.

Newman, D. J., Hettich, S., Blake, C. L., and Merz, C. J. (1998). UCI machine learning repository.

Nordhausen, K. and Ruiz-Gazen, A. (2022). On the usage of joint diagonalization in multivariate statistics. *Journal of Multivariate Analysis*, 188.

Peña, D., Prieto, F. J., and Viladomat, J. (2010). Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of Multivariate Analysis*, 101(9):1995–2007.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering method. *Journal of the American Statistical association*, 66(336):846–850.

Raymaekers, J. and Rousseeuw, P. J. (2022). Silhouettes and quasi residual plots for neural nets and tree-based classifiers. *Journal of Computational and Graphical Statistics*, pages 1–12.

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

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317.

Shirkhorshidi, A. S., Aghabozorgi, S., and Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 10(12).

Siddiqui, K. (2013). Heuristics for sample size determination in multivariate statistical techniques. *World Applied Sciences Journal*, 27(2):285–287.

Sinnott, R. O., Duan, H., and Sun, Y. (2016). *Chapter 15 - A Case Study in Big Data Analytics: Exploring Twitter Sentiment Analysis and the Weather*, pages 357–388. Morgan Kaufmann.

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

Steinhaus, H. (1956). Sur la division des corps materiels en parties. *Bull. Acad. Polon. Sci., C1. III vol IV*, pages 801–804.

Tyler, D. E., Critchley, F., Dumbeng, L., and Oja, H. (2009). Invariant co-ordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):549–592.

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

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774.

A Feature Description

Variable	Description
Comp	Compactness
Circ	Circularity
D.Circ	Distance Circularity
Rad.Ra	Radius ratio
Pr.Axis.Ra	pr.axis aspect ratio
Max.L.Ra	max.length aspect ratio
Scat.Ra	scatter ratio
Elong	elongatedness
Pr.Axis.Rect	pr.axis rectangularity
Max.L.Rect	max.length rectangularity
Sc.Var.Maxis	scaled variance along major axis
Sc.Var.maxis	scaled variance along minor axis
Ra.Gyr	scaled radius of gyration
Skew.Maxis	skewness about major axis
Skew.maxis	skewness about minor axis
Kurt.maxis	kurtosis about minor axis
Kurt.Maxis	kurtosis about major axis
Holl.Ra	hollows ratio
Class	type

Table 5: Feature description for $Vehicle$ dataset.

B Tables of Results

Tables contain average results over 100 simulations for each of the four cases. Median, minimum and maximum values can also be found here.

			K-means				EM			
	p	**Avg. Comp. Nr.**	**Avg. ARI**	**Min ARI**	**Max ARI**	**Median ARI**	**Avg. ARI**	**Min ARI**	**Max ARI**	**Median ARI**
Original Data	10	-	0.0562	0.0249	0.0844	0.0543	0.4921	0.4865	0.7273	0.4898
	50	-	0.0086	-0.0037	0.0239	0.0086	0.0315	-0.0124	0.4898	0.0115
	100	-	0.0037	-0.0067	0.0172	0.0032	0.0063	-0.0089	0.0323	0.0041
PCA	10	1.16	0.0798	0.0267	0.2583	0.0605	0.0584	-0.0107	0.3561	0.03
	50	2	0.0353	0.0174	0.0621	0.0345	0.03	-0.0233	0.0904	0.0284
	100	2	0.0164	0.0035	0.0339	0.0168	0.0149	-0.0298	0.064	0.0162
ICS1	10	1.55	0.3766	0	0.4648	0.3893	0.4096	0	0.4679	0.4196
	50	0.9	0.0049	-0.0106	0.0992	0	0.0015	-0.0373	0.0564	0
	100	1.42	0.0017	-0.0081	0.0244	0.0006	-0.0018	-0.0174	0.0194	-0.0019
ICS2	10	2	0.4048	-0.0409	0.4898	0.4898	0.4897	0.4856	0.4898	0.4898
	50	0.54	0.0009	-0.0065	0.0197	0	-0.0015	-0.0259	0.0065	0
	100	0.8	0.0011	-0.0091	0.0171	0	-0.0008	-0.0145	0.0235	0
ICS3	10	2	0.4078	-0.0409	0.4898	0.4898	0.4897	0.4856	0.4898	0.4898
	50	0.71	0.0022	-0.0065	0.0272	0	-0.0011	-0.0214	0.0244	0
	100	0.86	0.0009	-0.0079	0.0146	0	-0.0014	-0.0223	0.025	0

Table 6: Case 1 results. ICS1 is ICS with COV & COV_4, ICS2 is ICS with MCD & COV, ICS3 in this case is ICS with MCD $\alpha = 0.75$ & COV.

			K-means				EM			
	p	**Avg. Comp. Nr.**	**Avg. ARI**	**Min ARI**	**Max ARI**	**Median ARI**	**Avg. ARI**	**Min ARI**	**Max ARI**	**Median ARI**
Original Data	10	-	0.2298	0.1749	0.2893	0.2299	0.6301	0.4865	0.7279	0.4898
	50	-	0.0578	0.0311	0.0796	0.0565	0.3798	-0.0038	0.4898	0.0115
	100	-	0.0299	0.0129	0.056	0.0293	0.041	-0.0083	0.0328	0.0041
PCA	10	1.96	0.4071	0.2278	0.5331	0.3939	0.5865	0.2139	0.715	0.6454
	50	2	0.1791	0.1377	0.2342	0.1777	0.2059	0.0292	0.3745	0.2024
	100	2	0.1051	0.0715	0.1578	0.1035	0.0864	-0.0172	0.2061	0.085
ICS1	10	1.55	0.3766	0	0.4648	0.3893	0.4096	0	0.4679	0.4196
	50	0.9	0.0049	-0.0106	0.0992	0	0.0011	-0.0373	0.0566	0
	100	1.42	0.0017	-0.0081	0.0244	0.0006	-0.0017	-0.0174	0.0194	-0.0013
ICS2	10	2	0.4048	-0.0409	0.4898	0.4898	0.4897	0.4856	0.4898	0.4898
	50	0.54	0.0009	-0.0065	0.0198	0	-0.0015	-0.0259	0.0065	0
	100	0.8	0.0011	-0.0092	0.0171	0	-0.0008	-0.0145	0.0235	0
ICS3	10	2	0.4078	-0.0409	0.4898	0.4898	0.4897	0.4856	0.4898	0.4898
	50	0.71	0.0022	-0.0065	0.0272	0	-0.0011	-0.0214	0.0244	0
	100	0.86	0.0009	-0.0079	0.0146	0	-0.0014	-0.0223	0.025	0

Table 7: Case 2 results. ICS1 is ICS with COV & COV_4, ICS2 is ICS with MCD & COV, ICS3 in this case is ICS with MCD $\alpha = 0.75$ & COV.

			K-means				EM			
	p	**Avg. Comp. Nr.**	**Avg. ARI**	**Min ARI**	**Max ARI**	**Median ARI**	**Avg. ARI**	**Min ARI**	**Max ARI**	**Median ARI**
Original Data	10	-	0.0573	0.0249	0.0868	0.0574	0.671	0.6675	0.6711	0.6711
	50	-	0.0092	-0.0019	0.0272	0.0089	0.033	-0.0292	0.6711	0.0076
	100	-	0.0043	-0.0056	0.0169	0.0039	0.0079	-0.0062	0.0737	0.0061
PCA	10	1.28	0.1021	0.0282	0.267	0.0656	0.0899	-0.0208	0.4156	0.0339
	50	2	0.0372	0.0179	0.0593	0.0365	0.0317	-0.0128	0.0874	0.0319
	100	2	0.0176	0.006	0.0339	0.0171	0.0141	-0.019	0.0522	0.0137
ICS1	10	2	0.6272	-0.0259	0.6711	0.6682	0.6419	-0.0143	0.6711	0.6711
	50	1.27	0.0027	-0.0069	0.0703	0.0002	-0.0028	-0.0214	0.0204	-0.0009
	100	1.6	0.0024	-0.0102	0.0363	0.0009	-0.0009	-0.0261	0.0256	-0.0004
ICS2	10	2	0.5263	-0.0268	0.6711	0.6682	0.6349	-0.0175	0.6711	0.6711
	50	1.4	0.0194	-0.0267	0.3552	0.0046	0.0131	-0.0202	0.4326	-0.0164
	100	0.94	0.0011	-0.0099	0.0234	0	-0.0019	-0.0227	0.0143	0
ICS3	10	2	0.5568	-0.0268	0.6711	0.6688	0.6353	-0.0167	0.6711	0.6711
	50	1.48	0.0614	-0.0268	0.5691	0.0052	0.07168	-0.0202	0.5589	-0.0139
	100	1.16	0.0009	-0.0104	0.0159	0	-0.0023	-0.0232	0.01261	0

Table 8: Case 3 results. ICS1 is ICS with COV & COV_4, ICS2 is ICS with MCD & COV; ICS3 in this case is ICS with MCD $\alpha = 0.75$ & COV for $p = 10, 50$, and for $p = 100$ ICS3 uses *deterministic* MCD & COV.

			K-means				EM			
	p	**Avg. Comp. Nr.**	**Avg. ARI**	**Min ARI**	**Max ARI**	**Median ARI**	**Avg. ARI**	**Min ARI**	**Max ARI**	**Median ARI**
Original Data	10	-	0.2359	0.1917	0.2818	0.2348	0.701	0.6675	0.784	0.6711
	50	-	0.0612	0.0404	0.083	0.0609	0.4152	-0.0102	0.7644	0.6703
	100	-	0.0322	0.0122	0.059	0.0324	0.039	-0.004	0.1735	0.0298
PCA	10	2	0.4486	0.3781	0.5595	0.4373	0.653	0.4771	0.7836	0.6785
	50	2	0.1915	0.1527	0.2613	0.1897	0.2113	0.0567	0.3681	0.2169
	100	2	0.1135	0.0769	0.1668	0.1118	0.0941	0.0265	0.1964	0.0924
ICS1	10	2	0.6272	-0.0259	0.6711	0.6682	0.6418	-0.0143	0.6711	0.6711
	50	1.27	0.0026	-0.0069	0.0703	0.0002	-0.0028	-0.0214	0.0204	-0.0009
	100	1.6	0.0024	-0.0102	0.0363	0.0009	-0.0008	-0.0261	0.0237	-0.0004
ICS2	10	2	0.5263	-0.0268	0.6711	0.6682	0.6349	-0.0175	0.6711	0.6711
	50	1.4	0.0195	-0.0267	0.3552	0.0046	0.015	-0.0202	0.4326	-0.0163
	100	0.93	0.0012	-0.0099	0.0234	0	-0.0017	-0.0177	0.0134	0
ICS3	10	2	0.5568	-0.0268	0.6711	0.6688	0.6353	-0.0167	0.6711	0.6711
	50	1.52	0.0353	-0.0269	0.542	0.0042	0.0285	-0.0193	0.5303	-0.0164
	100	1.06	0.0013	-0.0106	0.0209	0	-0.0038	-0.0227	0.0146	0

Table 9: Case 4 results. ICS1 is ICS with COV & COV_4, ICS2 is ICS with MCD & COV; ICS3 in this case is ICS with MCD $\alpha = 0.75$ & COV for $p = 10$, and for $p = 50, 100$ ICS3 uses *deterministic* MCD & COV.