Erasmus University Rotterdam

Erasmus School of Economics

Tandem Clustering of Rating Scale Data with Invariant Coordinate Selection

Master Thesis

Econometrics and Management Science

Author: Douwe Rikken (428453)

Supervisor: dr. Aurore Archimbaud

Co-reader: drs. Jeffrey Durieux

March 27, 2023

The author of this thesis takes full responsibility for its content, which does not represent the views of the university, the supervisor, or the co-reader.

Abstract

Clustering of Rating Scale Data (RSD) is a challenging process that has been examined in previous research, with often inconclusive results. The issues of non-continuity, ordinality and dimensionality all have an impact on the ability of researchers to apply multivariate analysis, for example clustering algorithms such as $K-means$. Tandem clustering could be an interesting new approach. When tandem clustering is used, it is assumed that the important structure of the data is on a lower dimensional subspace. To address this issue, in tandem clustering a dimension reduction method is firstly applied, before the clustering algorithm. Usually tandem clustering is used with Principal Component analysis to deal with the dimension reduction. However, a new method has been introduced. In this context Invariant Coordinate Selection (ICS) is used as a method for dimension reduction. To adequately apply ICS to perform dimension reduction, a few challenges need to be dealt with. Firstly, the choice of the appropriate scatter matrices is crucial to discover the structure on the Invariant Coordinates. Secondly, the selection of those Invariant Coordinates, that actually display the relevant structure, is essential. Following the application of the ICS method, the lower dimensional data is spanned by the selected "Invariant Coordinates". The observations of the Invariant Coordinates are then clustered by the $K-means$ algorithm. In an attempt to broaden the use of the method, in this thesis, ICS is applied to RSD and subsequently clustering is performed through $K-means$. Then, the performance of the method is reviewed by calculating the Adjusted Rand Index (ARI). Of the chosen scatter pairs, $TM-COV$ and $COV-COV4$ perform relatively well. The choice of the scatter matrices is not optimal and different options should definitely be explored. Additionally, the selection of the Invariant Components presents issues. An automatic selection process is required to make the method more accessible, however, the automatic selection criteria presented in this research, the D'Agostino test and the Med criterion, do not perform well enough. To deal with this issue, in this research the relevant Invariant Coordinates are selected manually. To improve performance of the tandem clustering algorithm, the option of taking the average of the constructs within the data is explored, to extend tandem clustering with ICS to deal with RSD. Averaging the constructs consists of taking the mean over the Likert items, within the constructs of the data, of every observation. Since this transformation makes the data closer to continuous it is expected that this will improve performance, which is actually what occurs. Generally, the results justify further research into the relevant choice of scatter pairs for ICS when it is applied to RSD, as well as into the method for component selection.

Contents

1 Introduction

From customer satisfaction surveys to clinical trials, Rating Scale Data (RSD) has become an increasingly popular and essential tool for measuring subjective experiences and opinions in various fields of research and application. RSD is a type of data that consists of a scale of generally 5 or 7 points, that are in increasing order and have labels attached to them. These labels usually describe the measure of a person's agreement to a particular statement.

The Likert Scale is a type of RSD introduced by Likert (1932). The scale generally consists of a 5 or 7 point ordinal or interval line, representing ranked answers to a given question or statement. Likert Scale data is often obtained through surveys or questionnaires consisting of many statements or questions. Each of these statements is called a Likert item. Table 1 gives two commonly used examples of possible answers a Likert Scale could represent.

Likert Rating Scales				
1	2	3	4	5
never strongly disagree	rarely somewhat disagree	sometimes neutral	often somewhat agree	always strongly agree

Table 1: Likert scales, a form of RSD that attaches numbers to verbal answers to a question or statement.

The two examples given above are ordinal scales. Ordinal meaning the variables are ordered from, for example, worst to best or from least frequent to most frequent. However, the Likert Scale is generally interpreted as an interval scale (Wu and Leung, 2017), where the variables are thus not only ordered, but also equidistant from each other. Now even though this is not necessarily true for the two scales in Table 1, it is still often assumed to be true for comparable scales (Blaikie, 2003).

There is much discussion about how RSD should be analyzed. As described by Harpe (2015), possible issues include the controversy about whether this type of data is ordinal or continuous, and related to that, whether parametric or non-parametric approaches should be used when analyzing this data. But, the analysis of RSD in general is a field of research in which a lot of things are unclear. As Harpe (2015) indicates, even the most fundamental properties of this type of data cannot be agreed upon by researchers. Applying a clustering algorithm to RSD presents one of these many challenges. The discrete nature of the data makes it difficult to utilize algorithms such as $K - means$ clustering, because the calculation of Euclidean distance is not meaningful when dealing with categorical data. Categorical data do not have a natural origin and therefore the distance measure does not carry the same meaning. Multiple approaches have been introduced (Podani, 2005; Walesiak and Dudek, 2010; Jacques and Biernacki, 2018) to apply clustering analysis on RSD. However, assuming that the important structure of the discrete data is on a lower dimensional subspace, dimension reduction combined with clustering (tandem clustering) applied on RSD could be an interesting new approach. Tandem clustering was first introduced by (Arabie, 1994) in conjunction with Principal Component Analysis (PCA). The method works by selecting the relevant principal components and hereby performing dimension reduction. Subsequently, the clustering algorithm is applied to the relevant, lower dimensional data.

The Invariant Coordinate Selection (ICS) method, described by Tyler et al. (2009), is a method for examining multivariate data, that might be useful when analyzing RSD through tandem clustering. The method compares the eigenvalue-eigenvector decomposition of one scatter matrix relative to another. ICS has several (extra) attractive properties compared to other methods such as PCA. These properties allow ICS to be more efficient in finding outliers or clusters in the data, or to be not scale dependent contrary to PCA. Even though ICS has these advantages over other methods, it is not widely applied or studied in clustering applications for example. This is partly explained by it being a relatively new method. Also, its application on Rating Scale Data has not been studied. Furthermore, it was concluded by Fischer et al. (2017) that ICS is able to recover lost subgroups within simulated as well as real world data, while PCA was unable to detect any subgroups. This gives a clear indication that further studies of the application of ICS are warranted, especially when the widespread use of PCA is considered. The method has been succesfully applied in a tandem clustering context by Alfons et al. (2022). However, this was an application on continuous data. It is thus far unclear whether its application is justified in a rating scale context. The research question will therefore be:

Can ICS be used to perform tandem clustering when applied to Rating Scale Data? How does it compare to PCA in this context? What extensions can be made to the ICS method in order to make it more applicable to this type of data?

This research will formulate a comprehensive answer to this question in the following way: Firstly, the RSD set will be simulated in R (R Core Team, 2020) using the models suggested by Bernaards and Sijtsma (2005) and Wu and Leung (2017). The objective is to make them contain multiple high complexity clusters. This method of generating data will be an essential part of this research as it is critical that this is done in a proper way, to make sure that the ICS method can then be appropriately applied. Furthermore, the general nature of RSD will be analyzed using the existing literature. Secondly, the application and possible extension of the ICS method to a tandem clustering approach in a Rating Scale Data context. The goal is to use ICS for dimension reduction and to perform clustering and compare its performance relative to PCA. The application of ICS on RSD, as well as its clustering performance are two new fields of research which could increase the utility of the ICS method even more. In general, the limitations of the methods introduced in Alfons et al. (2022) will be shown when the method is applied in a similar way to RSD. These limitations will then be analyzed and a possible solution will be tested. For example, Harpe (2015) provides insight to that problem and gives an indication of the criteria that should be met for ICS to be appropriately applied on rating scale data. The thesis will explore the possibility of extending ICS to this specific type of data to make it more applicable and useful in further studies.

This paper will present an overview of the literature on the clustering of RSD in Section 2. Section 3 explains the Methodology with regards to the method of Invariant Coordinate Selection as well as the clustering method. The setup of the simulation will be introduced in Section 4. Finally, the results will be discussed in Section 5.

2 Literature Review

2.1 Challenges with RSD analysis

The general conviction when examining a Likert Scale was that the variables are ordinal but not equidistant and thus not on an interval scale. Therefore, only non-parametric techniques should be used when analyzing this type of data. This consensus could have originated from Stevens (1951), who stated that means and standard deviations, some of the anchors of parametric statistics, were not suitable metrics of central tendency for ordinal data (Pett, 2015). A frequently used argument for this point of view is, as stated by Jamieson (2004, p.1218), "the average of fair and good is not fair-and-ahalf". Which is exactly the argument proposed by Sullivan and Artino Jr (2013) and Kuzon et al. (1996). This is generally a convincing argument because it does not make sense to take an average of words, or perform any mathematical operation on data obtained from words. The fact that the Likert Scale proposes to label those words with numbers does not change that. However, it should not matter whether or not it makes sense to people if we perform those operations on this data, what is relevant is whether or not the chosen statistics will lead to the right conclusion. Specifically, if the use of parametric methods leads to the correct conclusions, even when it violates certain assumptions (in this case the assumption of continuity of the ordinal/interval data), then the use of it is justified (Norman, 2010; Sullivan and Artino Jr, 2013).

There are several other convincing arguments for the use of parametric methods on ordinal data (Pell, 2005; Carifio and Perla, 2007, 2008; Norman, 2010). For example, Knapp (1990) argued that if the data is relatively normally distributed and of an acceptably large sample size ($N > 30$), then the ordinal nature of the data is not as much of an issue and the use of parametric methods could be justified. Additionally, Norman (2010) made a reasonable claim that parametric statistics are of good use on Likert Scale data even when assumptions regarding normality, sufficiently large sample size and equal variances are violated. His argument that the violations of the sample size and normality assumptions are not an issue, is that parametric methods, such as the Analysis Of Variance (ANOVA), can deal with skewness and non-normality, because they are highly robust to them (Norman, 2010). However this argument does not cover the issue of unequal variances. These present an issue when dealing with correlations and regression coefficients. However, convincing research conducted by Pearson (1931, 1932a,b) and Havlicek and Peterson (1976) concludes that the Pearson correlation is exceptionally robust against the violation of the equal variance assumption.

Another remark is made by Wu and Leung (2017), who argue that increasing the Likert Scale to a scale of 11 points would better allow for its use as an interval scale. Specifically, increasing the scale to 11 points contributes to a closer approach of the underlying distribution. Leung (2011) confirmed this by finding that increasing the number of points on the scale does not lead to a change in the means and standard deviations, but reduced the skewness and kurtosis. Thus, this results in a closer approach to a normal distribution and therefore an interval scale.

The choice between considering Likert data as ordinal or interval is obviously of high importance. As indicated, interval data allows for the use of parametric methods which have several advantages over non-parametric methods. For example: the use of more powerful, sensitive and better interpretable statistics, retaining more information about the nature of the data, more options in manipulating the data (regression analysis, analysis of variance and covariance etc.) (Labovitz, 1970).

2.2 Clustering Rating Scale Data

The application of clustering algorithms to RSD has been a frequent topic of research. Walesiak and Dudek (2010) evaluated a number of different clustering procedures, consisting of 9 different clustering algorithms and, for each algorithm, 8 different within cluster quality indices. These within cluster quality indices are used to determine the number of clusters and could be the Davies-Bouldin (Davies and Bouldin, 1979) or the Calinski-Harabasz (Caliński and Harabasz, 1974) index for example. A ranking for the procedures was formed based on the Adjusted Rand Index (ARI)[*]. The index compares the clustering to the true clustering of

[*] The Adjusted Rand Index is a measure that gives an indication of the affinity of two clusterings. A clustering being the allocation of the data into the relevant amount of clusters. If the two clusterings are

the data, which is known because the data is simulated. The best performing algorithm was group average link, while the worst performing method was the single-link algorithm. In the group average link algorithm, the linkage function specifies the distance between two clusters as the average of all pairwise distances of the points in each cluster. This makes group average link more robust to outliers. In the single-link algorithm, the linkage function specifies the distance between two clusters as the shortest distance possible between two points, one in each cluster. This makes single linkage more sensitive to outliers and therefore less robust.

Jacques and Biernacki (2018) performed a co-clustering algorithm on ordinal data and also evaluated the clustering result with the Rand Index. Co-clustering is a method that involves creating a pair of mappings, one from rows to clusters of rows, and another from columns to clusters of columns (Dhillon et al., 2003). Jacques and Biernacki (2018) propose two simulation settings where the first contains clusters that are well separated, whereas the second contains clusters that are more mixed. The Adjusted Rand Index for the first setting was 0.97 for both the rows and the columns, while the Adjusted Rand Index for the second setting was 0.58. It has to be noted that both had a relatively high standard deviation of $\sigma = 0.13$ and $\sigma = 0.16$ respectively[†].

Van de Velden et al. (2017) propose a method for categorical variables that simultaneously provides an optimal split of the clusters as well as a dimension reduction through the association of the so-called "active variables". They call this method "cluster correspondence analysis". For a deeper understanding of their simulation and dimension reduction we recommend reading Van de Velden et al. (2017). For their balanced cluster scenario (four equal-sized clusters), and for their proposed method, they achieve a minimal Adjusted Rand Index of 0.08 in a certain setup for dimension reduction, and a maximum Rand Index of 0.99 for a different setup. The achieved Rand Index seems to vary quite intensely based on how many active variables are selected, as well as the chosen association strength between the active variables and the clusters, the Cramér's V. The Cramér's V is a measure of the

identical, i.e. each cluster contains the exact same points, the value of the Rand index is equal to 1, while the Rand index is equal to 0 if the clusterings have no affinity at all, i.e. contain no common points.

[†] Aggregated over the row and column Indices

affiliation among variables and was first introduced by Cramér (1999). Van de Velden et al. (2017) use two different values, 0.5 and 0.7. The results are significantly better for a higher Cramér's V as well as for a higher number of active variables. This seems to indicate that their dimension reduction could be less impressive than reduction done by ICS. The number of total variables differs for the low and high noise scenarios. For low noise, the number of active variables is added as noise variables, meaning variables that have no affinity to the cluster. Therefore, for the low noise scenario, the dimension of the data is only halved. For the high noise scenario, which achieved worse performance on average, four times the number of active variables is added as noise. In general, the number of active variables needed to assign the clusters relatively well seems to be significantly higher than the number of clusters.

In general, these methods do not seem to solve the issue of finding the relevant structure on the lower dimensional subspace. Except the methods introduced by Van de Velden et al. (2017), those are dealing with dimension reduction and show promising results. This is where tandem clustering with ICS can be helpful.

3 Methodology

As discussed in Section 1, tandem clustering is a procedure that combines a method for dimension reduction with a clustering algorithm. It can be applied when the assumption is made that the relevant structure of the data lies on a lower dimensional subspace. A method like PCA or ICS is then applied first to find that lower dimensional subspace, in order to then apply the clustering algorithm to this lower dimensional data. This is then supposed to improve on the original clustering algorithm without dimension reduction.

3.1 Invariant Coordinate Selection

The use of a method for Invariant Coordinate Selection was first introduced by Tyler et al. (2009). The method is based on the comparison of the eigenvalue-eigenvector decomposition of two affine equivariant scatter matrices. A scatter matrix is a statistic used for the robust estimation of a covariance matrix when that particular matrix is considered to be biased or inconsistent because of the nature of the data. A scatter matrix is affine equivariant if

$$S(X_nF + 1_nb') = F'S(X_n)F, \tag{1}$$

where X_n is an $n \times p$ matrix, F is a full rank $p \times p$ matrix, b is a p-dimensional vector and 1_n is an n-dimensional vector containing only ones. There are different classes of scatter matrices introduced by Tyler et al. (2009), each with different robustness properties. Their robustness can be measured by a breakdown point and an influence function. An estimator with a high breakdown point and a bounded influence function is considered robust, as opposed to an estimator with a breakdown point of zero and an unbounded influence function, which is not robust at all. Specifically, Tyler et al. (2009) show that an affine invariant coordinate system can be introduced based on the eigenvectors of these scatters. Subsequently, this system helps to display distinctive structures within the data. Furthermore, ICS is able to recover Fisher's linear discriminant subspace Tyler et al. (2009) even if the class identifiers are unknown. The linear discriminant was first conceptualized by Fisher (1936) and further generalized to a subspace by Rao (1948). The general idea of ICS can be illustrated as in the equations:

$$\begin{aligned} B(X_n)S_1(X_n)B(X_n)' &= I_p, \\ B(X_n)S_2(X_n)B(X_n)' &= D(X_n), \end{aligned} \tag{2}$$

as suggested in Archimbaud et al. (2018). Where, $S_1(X_n)$ and $S_2(X_n)$ are the scatter matrices and $D(X_n)$ is a diagonal matrix displaying the eigenvalues of the matrix $S_1(X_n)^{-1}S_2(X_n)$ in decreasing order. The rows of $B(X_n)$ contain the corresponding eigenvectors. If $\mu(X_n)$ is an affine equivariant location estimator then the invariant components are given in

$$IC_n = (ic_1, \ldots, ic_n)' = (X_n - 1_n\mu(X_n)')B(X_n)'. \tag{3}$$

ICS finds the $B(X_n)$ and $D(X_n)$ matrices such that the equations in 3 hold. In order to do this, firstly, the choice of a pair of scatter matrices must be made. Secondly, the relevant Invariant Coordinates need to be selected. These two tasks form the main challenges with ICS and are presented in the next Sections.

3.1.1 Choice of Robust Scatter Matrices

Because of possible non-normality of the data and possible outliers, it is essential that robust estimates of multivariate location and scatter are used. The multivariate M-estimates introduced by Maronna (1976) and Huber (1981) are examples of a robust estimator. They are detailed as a class 2 estimator by Tyler et al. (2009) meaning that they are somewhat robust, referring to a positive breakdown point, but have a breakdown point limited by $1/(p-1)$. Class 1 and class 3 estimators are given by, for example, the sample covariance matrix and the S-estimates respectively (Davies, 1987; Lopuhaa, 1989). Class 1 estimators are characterized by not being robust and a breakdown point equal to zero. Class 3 estimators are thus characterized by a very high breakdown point.

Class 1 estimators are indicated to be useful in a situation where the data set does not contain any outliers (Caussinus and Ruiz-Gazen, 1990), or if the explicit goal is to detect those outliers. However, existing research indicates that choosing robust estimates for the ICS transformation can improve performance (Nordhausen et al., 2008a). It is therefore recommended to use class 2 or class 3 estimators, because they are still able to find outliers using the Mahalanobis distances, but they are not heavily affected by those outliers in an ICS transformation (Tyler et al., 2009). Class 2 estimators are relatively more affected by outliers than class 3 estimators, especially when those outliers form a cluster (Dümbgen and Tyler, 2005). A potential drawback of class 2 estimators is their low breakdown point, especially when handling high-dimensional data. Class 3 estimators do not have this problem, however they do have the issue of computational intensity. This is a more significant problem for high dimension and large data sets.

Tyler et al. (2009) give a recommendation to use different combinations of one class 2 and one class 3 estimate of scatter as a general choice, while Archimbaud et al. (2018) propose the use of two class 1 matrices. The covariance matrix (COV) and the scatter matrix of fourth moments (COV_4). This general choice is motivated by the simplicity and relative ease of computation of these estimators. Also, Archimbaud et al. (2018) argue that the Theorems

3 and 4 given in Tyler et al. (2009) still hold for these estimators, even though they are not robust. However, this recommendation was made with a prerequisite: Caussinus and Ruiz-Gazen (1990), Caussinus et al. (2003) and Tyler et al. (2009) recommend using Class I scatter estimators, if the objective is outlier detection (Archimbaud et al., 2018). It is therefore not clear whether the choice of two class 1 scatter matrices is appropriate for other objectives such as clustering purposes.

An example of class 1 estimators, are the sample means and sample covariance matrices. The multivariate M-estimates proposed by Maronna (1976); Huber (1981) are an example of an estimator of class 2, given by,

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{n} u_1(s_i) y_i \Big/ \sum_{i=1}^{n} u_1(s_i), \\ \hat{V} &= \sum_{i=1}^{n} u_2(s_i)(y_i - \hat{\mu})(y_i - \hat{\mu})' \Big/ \sum_{i=1}^{n} u_3(s_i), \end{aligned} \tag{4}$$

in Tyler et al. (2009). Where $s_i = (y_i - \hat{\mu})'\hat{V}^{-1}(y_i - \hat{\mu})$, and $u_1(s)$,$u_2(s)$ and $u_3(s)$ are some chosen weight functions. In the ICS package (Nordhausen et al., 2008b) in R, the function tM is provided which is a variant of the M-estimator in equation 4. The tM estimator uses three EM-algorithms described in Kent et al. (1994) to make an M-estimation of the multivariate location and scatter of a t-distribution. The tM function is a good and robust alternative for the M-estimator. Because it is based on the t-distribution it allows for heavier tails and is therefore less sensitive to outliers.

Finally, the reweighted Minimum Covariance Determinant (MCD) estimator is an example of a class 3 estimator, as suggested in Archimbaud et al. (2018). Cator and Lopuhaä (2012) defined the MCD as

$$\begin{aligned} \hat{\mu}_n(S) &= \frac{1}{h_n} \sum_{X_i \in S} X_i \\ \hat{V}_n(S) &= \frac{1}{h_n} \sum_{X_i \in S} (X_i - \hat{\mu}_n(S))(X_i - \hat{\mu}_n(S))', \end{aligned} \tag{5}$$

for the sample $X_1, X_2, \ldots, X_n$ and $0 < \gamma < 1$ and the subsamples $S \subset [X_1, X_2, \ldots, X_n]$ containing $h_n \geq \lceil n\gamma \rceil$ points (meaning $\lceil n\gamma \rceil$ is the smallest integer $\geq n\gamma$). If S_n is the subsample that minimizes $det(\hat{V}_n(S))$ over all the subsamples of size $h_n \geq \lceil n\gamma \rceil$, then $(\hat{\mu}_n(S_n), \hat{V}_n(S_n))$ is the MCD estimator. The α parameter in the MCD estimator specifies what fraction of the data should be used to calculate the Scatter matrix. With regards to ICS it is suspected by Alfons et al. (2022) that a low α value might be useful because it could allow the MCD estimator to capture within cluster structure. However, α is chosen to be equal to 0.6 because of computational issues that arise, especially in lower dimensions. These issues arise because of the discrete nature of the data.

Finally, the $LCOV_{S_0,\beta}$ estimator as it was introduced in Alfons et al. (2022), is an estimator with a specific application in an ICS context. Like the MCD estimator, it is able to capture the within cluster variance, while the other scatter matrix might be able to capture the overall structure. This estimator was first suggested by Hennig (2009) and can be interpreted as a local shape matrix. S_0 is a scatter matrix (COV will be used) and β is a proportion that indicates how many nearest neighbors ($n_\beta = \lceil \beta n \rceil$) should be used to calculate the COV matrix for each observation. Subsequently, these covariance matrices are averaged after scaling.

As discussed, in ICS it is often suggested that several combinations of different class scatter matrices (the so-called “scatter pairs”) are used, because of the lack of theory on which combination of scatter matrices works best in new scenarios (Tyler et al., 2009)(Nordhausen et al., 2008b). Since Alfons et al. (2022) performed an extensive data simulation with the purpose of tandem clustering with ICS, it is useful to examine the scatter pairs used in their research. Their selection of scatter pairs included $LCOV - COV$, $MCD_\alpha - COV$, $MCD_{0.25} - MCD_{0.95}$ and $COV - COV4$. In Section 4.5, the chosen pairs for this research will be further explained and justified.

3.1.2 Selection of Invariant Coordinates

As Fischer et al. (2017) indicated, clusters are often found in the directions of extreme (low or high) kurtosis values. More specifically, large clusters correspond to low kurtosis values and small clusters to high kurtosis values. The Invariant Coordinates (IC) in ICS are therefore especially useful since they are ordered according to their kurtosis. Thus, the coordinates indicate in which direction clusters can be detected. However, as Alfons et al. (2022) indicated, the direction that clusters can be detected will strongly depend on several factors such as choice of scatter matrices, the underlying distribution and cluster structure. Therefore, it cannot be assumed that clusters will always be found in a particular direction.

When performing dimension reduction, it is important to choose specific dimensions such that the important structure and variance in the data is preserved. When selecting the invariant coordinates, one way of detecting which coordinates capture the structure, is to check if the specific coordinates significantly deviate from the normal distribution. To achieve this, the D'Agostino test will be used. This is a skewness test that checks if the distribution of the data is significantly different from a normal distribution. If that is the case for a specific component, it will be selected. However, since we know from previous theory that the structure can be found on a subset of the first and last components (Alfons et al., 2022), the algorithm for selecting this subset is not as straightforward. The algorithm iteratively checks whether the first component is significantly different from normal at a level of $\alpha = 0.05$. If that is the case, then it checks whether the last component is significant at an adjusted level of $\alpha = 0.025$. This continues, until a component is found that is not significant. If for example, this component is found at the front, then subsequently only components from the back will be checked and vice versa. This will continue until one component from both sides has been found to be insignificantly different from a normal distribution.

Another method that will be utilised is the *med criterion*, introduced by Alfons et al. (2022). This criterion compares the eigenvalues of the Invariant Coordinates with their mean and selects the $q - 1$ Coordinates whose eigenvalues is furthest away from the mean. Here q is equal to the number of clusters in the data. A different selection method like this one is required because of the D'Agostino test's inability to detect any skewness if the clusters are distributed in a $50-50$ setup. Meaning, the two clusters will cancel eachother out and the test will fail to detect skewness.

3.2 Tandem Clustering

Alfons et al. (2022) research the option of tandem clustering using ICS. Tandem clustering being the process of first applying a method for dimension reduction before applying the clustering algorithm. The process of dimension reduction is performed by ICS as it was described in Section 3.1. Alfons et al. (2022) conclude that for the purpose of clustering the best choice of scatter matrices is made by selecting two scatters, such that one conveys the within-cluster structure and another the overall structure. Furthermore, they focus on the specific scatter pairs of $LCOV-COV$, $TCOV-COV$, $TCOV-UCOV$ and $COV-COV4$, in combination with the *med criterion*, as well as $MCD_\alpha - COV$ for the D'Agostino test criterion. Furthermore, in the existing literature on ICS, usually the $K-means$ clustering algorithm is used. In this research, $K-means$ is the only clustering algorithm that will be considered, because the purpose of the research is not to compare performance between clustering algorithms. Instead, the purpose is to compare performance between dimension reduction methods (PCA and ICS) as well as different settings within the ICS method.

3.2.1 K-means Clustering

A clustering problem consists of having an intitial $n \times p$ data set $X_n = (x_1, \dots, x_n)'$ that has to be partitioned into K clusters $C_1, \dots, C_K$ such that a specific criterion is minimized. This clustering criterion is often chosen to be the sum of the squared distances between the observations x_i and the randomly chosen centers m_k(Likas et al., 2003). This clustering error can then be shown to be

$$CE(m_1, \dots, m_K) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k)|x_i - m_k|^2, \tag{6}$$

where $I(x_i \in C_k)$ is an indicator function that is equal to 1 if X is true and 0 otherwise.

The $K-means$ algorithm starts at arbitrary chosen centers m_k and finds local optima when minimizing the CE. It proceeds by placing the centers in a better position with each step in order to decrease the CE. A drawback is therefore that the $K-means$ algorithm is sensitive to local optima. Because of that, the $kmeans()$ function from the $stats$ package in R (R Core Team, 2020) is used to apply the clustering method. It is chosen to use twenty different starting points in order to prevent ending up in a local optimum. This is done by providing $nstart = 20$ to the function. Also the number of clusters $k = 1, 2, 3$ is provided up front, since we know the true clusters. To be able to objectively compare the performance of the two methods, an exact measure is required. As suggested by Fischer et al. (2017), the (Adjusted) Rand Index (Rand, 1971) can provide a solution here.

3.2.2 The Rand Index

As indicated in Section 1, the Rand Index will be used for performance evaluation. The Rand Index is a measure that was introduced by Rand (1971). The measure gives an indication of the affinity of two clusterings. A clustering being the allocation of the data into the relevant amount of clusters. If the two clusterings are identical, i.e. each cluster contains the exact same points, the value of the Rand index is equal to 1, while the Rand index is equal to 0 if the clusterings have no affinity at all, i.e. contain no common points. For N points, $Q_1, Q_2, \cdots, Q_N$, the Rand Index is given by

$$RI(X,Y) = \frac{\left[\binom{N}{2} - [1/2\sum_i(\sum_j n_{ij})^2 + \sum_j(\sum_i n_{ij})^2 - \sum\sum n_{ij}^2]\right]}{\binom{N}{2}}, \qquad (7)$$

as proposed by Rand (1971). Where in equation 7, $X = X_1, \cdots, X_{K1}$ and $Y = Y_1, \cdots, Y_{K2}$ are two clusterings of the N points. The clusters are numbered arbitrarily and indicated by subscripts i and j. Moreover, n_{ij} is defined as the number of points simultaneously in the ith cluster of X and the jth cluster of Y. Nowadays, the use of the Adjusted Rand Index, is much more common as it corrects for the random grouping of elements. It was introduced by Hubert and Arabie (1985) and is of the form

$$ARI(X,Y) = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]/\binom{n}{2}}. \tag{8}$$

where n_{ij},$n_{i\cdot}$ and $n_{\cdot j}$ are values from the clustering table. The clustering table is given by

Clustering Table					
Y X	Y_1	Y_2	$\cdots$	Y_s	sum
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
sum	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	

Table 2: Clustering table in relation to equation 8

The Adjusted Rand Index obviously rates the performance of one clustering method relative to another method. As Rand (1971) indicated, the evaluation of the performance of a clustering method requires a way to compare its results to the actual clustering, or to another method. For simulated data, where the true clusterings are known, the performance of one method can therefore be evaluated relative to the true clusterings. For the Adjusted Rand Index, a result of 1 means the performed clustering is exactly equal to the true clustering. A result of 0 means the performed clustering is not performing better than random chance would. A negative result would indicate that the clustering method is performing worse than random chance.

4 Simulation Design

4.1 Context of RSD Simulation

The popularity of RSD has made it widely available. However, in this research the simulation of the data is preferred, because the underlying distribution of the data needs to be predetermined and known, in order to compare the performance of the methods to the actual structure of the data. To simulate RSD there are various choices to be made. Firstly, an underlying distribution needs to be chosen and subsequently, a way to discretize the data, since a Likert Scale is obviously discrete. Wu and Leung (2017) introduce multiple options and explore all of them for their research. For the underlying distributions they consider two options, a very skewed gamma distribution and a symmetric normal distribution. Subsequently, for the discretization approach, they examine equal probability, equal interval width-, symmetric-bell-shape- and skewed discretization. This gives them six different conditions to simulate data from, which is relevant to their research since they are investigating the relevance of increasing the Likert Scale to 11 points.

Bernaards and Sijtsma (2005) suggest three different models to simulate RSD, each with different attractive properties: The Multidimensional Polytomous Latent Trait Model (MPLT), the Normal Ogive Item Response Theory Model (IRT) and the Discretized Normal model. One of these, the MPLT model, will be given in the Appendix. This model and the IRT and Discretized Normal models, allow for very specific customization of responses for each individual, the MPLT and IRT models especially, which makes them notably suitable for the generation of RSD. Compared to the models introduced by Wu and Leung (2017), these methods are more advanced and more capable of capturing real world RSD.

Additionally, Goldfeld and Wujciak-Jens (2020) introduced an R package (*simstudy*) (R Core Team, 2020) for simulating large-scale RSD with several options for the generation of cluster samples through the customization of answer probability- and correlation matrices. The *simstudy* package is a tool that allows for the simulation of data sets in a big scale survey context. It is designed to make it easy to create data sets that are customized to specific conditions. It also enables the user to modify the level of clustering within categorical ordinal data sets. The *genOrdcat* function (Goldfeld and Wujciak-Jens, 2020) is a function within the *simstudy* package for generating categorical ordinal data. While using this function the number of categories and the probability of each category can be easily specified. This can then be used to generate data sets with predetermined levels of clustering. Additionally, the *genOrdcat* function enables the specification of the prevalence of each category, providing further control over the level of clustering in the data set. Therefore, the *simstudy* package allows user to easily simulate (categorical) data sets with predetermined levels of clustering.

Moreover, the *lsasim* package is an R package introduced by Matta et al. (2018) and designed for the simulation of complex data structures also in the context of big scale surveys. It is a useful package for simulating ordinal Likert scale data. The package allows for the generation of data with personally defined characteristics, such as nominal, ordinal, and continuous distributions. It is also possible to specify correlations between variables, and to generate data with block structure correlations. The *lsasim* package is a valuable package for data simulation, especially for those simulating RSD.

The data simulation in this study will be done in the context of big scale data surveys. The questions will be posed around a few different topics as is common with Likert scale data surveys (Nemoto and Beglar, 2014). Items that relate to a specific topic form a group or 'construct'. These constructs will be simulated by the blocks in the correlation matrix, which will be assessed in Section 4.4.

4.2 Simulation Setup

In R (R Core Team, 2020), the data will be simulated using the *simstudy* package. The *simstudy* package is very useful for the simulation of RSD, especially relating to the approach taken by Wu and Leung (2017). The choice for the *simstudy* package over the *lsasim* package is made mainly due to personal preference. The *simstudy* package offers the required options for customizing the data, such as specifying probabilities for different answer options or correlations between variables. However, the *lsasim* package also has these capabilities.

A thousand iterations of the simulation will be conducted, with each iteration generating data for 1000 individuals on three scenarios. The three scenarios will have 10, 25 and 50 Likert Scale items respectively. This seems to be a good variety of dimensions for the data to allow for sufficiently complex clusters and dimension reduction. Furthermore, doing a thousand iterations of the simulation will allow for the drawing of stronger and more robust conclusions with respect to the performance of the ICS method. Scenario 1, Scenario 2 and Scenario 3, with 10, 25 and 50 Likert items respectively, will each be simulated with no clusters, two clusters and three clusters. This is illustrated in Table 3 below.

Simulation Scenarios			
Scenario	Number of Likert Items	Number of Clusters	Rating Scale Length
1.1	10	0	5, 7 and 11
1.2	10	2	5, 7 and 11
1.3	10	3	5, 7 and 11
2.1	25	0	5, 7 and 11
2.2	25	2	5, 7 and 11
2.3	25	3	5, 7 and 11
3.1	50	0	5, 7 and 11
3.2	50	2	5, 7 and 11
3.3	50	3	5, 7 and 11

Table 3: All the scenarios that will be simulated and their corresponding properties.

The rating scale length mentioned in the table refers to the different formats of the Rating Scale that will be considered. Because the non-continuous nature of the data might be an issue, increasing the length of the rating scale could be helpful as mentioned in Section 2.1.

In the Scenarios that contain clusters, those clusters will be constructed as follows. In scenario 2.2, the 2 clusters will each represent 50% of the population. These two clusters will consist of a group that has no contamination in their answer probabilities to all questions, and a group that has the answers to questions 1 to 5 contaminated. In Scenario 2.3, the 3 clusters will each represent 33% of the population and will consist of groups that have their answers contaminated to questions 1-5, questions 11-15 and questions 21-25 respectively.

The setup for Scenario 3 is quite similar and both Scenarios are more clearly illustrated in the table below:

Clustering setup						
Scenario	Clusters 1	Size	Cluster 2	Size	Cluster 3	Size
1.2	No contamination	50%	Q01 - Q02	50%	-	-
1.3	Q01 - Q02	33%	Q04-Q06	33%	Q09-Q10	33%
2.2	No contamination	50%	Q01 - Q05	50%	-	-
2.3	Q01 - Q05	33%	Q11-Q15	33%	Q21-Q25	33%
3.2	No contamination	50%	Q01 - Q10	50%	-	-
3.3	Q01 - Q10	33%	Q21-Q30	33%	Q41-Q50	33%

Table 4: Every scenario and its subscenarios with their corresponding cluster setup. All setups portray a balanced clustering. Meaning all clusters are of the same size.

4.3 Answer probability matrices

The answer probability vectors, for the non-contaminated as well as the contaminated answers, for the 5-point, 7-point and 11-point Likert scale are shown in Table 5 below:

Answer Probability Vectors												
Likert Item Likert Scale	1	2	3	4	5	6	7	8	9	10	11	Contaminated
5-point centered	0.05	0.20	0.50	0.20	0.05	-	-	-	-	-	-	no
5-point shifted left	0.20	0.50	0.20	0.05	0.05	-	-	-	-	-	-	no
5-point shifted right	0.05	0.05	0.20	0.50	0.20	-	-	-	-	-	-	no
5-point contaminated	0.50	0.30	0.15	0.04	0.01	-	-	-	-	-	-	yes
7-point centered	0.04	0.11	0.20	0.30	0.20	0.11	0.04	-	-	-	-	no
7-point shifted left	0.11	0.20	0.30	0.20	0.11	0.04	0.04	-	-	-	-	no
7-point shifted right	0.04	0.04	0.11	0.20	0.30	0.20	0.11	-	-	-	-	no
7-point contaminated	0.45	0.25	0.15	0.055	0.04	0.03	0.025	-	-	-	-	yes
11-point centered	0.02	0.03	0.05	0.08	0.17	0.30	0.17	0.08	0.05	0.03	0.02	no
11-point shifted left	0.03	0.05	0.08	0.17	0.30	0.17	0.08	0.05	0.03	0.02	0.02	no
11-point shifted right	0.02	0.02	0.03	0.05	0.08	0.17	0.30	0.17	0.08	0.05	0.03	no
11-point contaminated	0.4	0.2	0.12	0.08	0.05	0.036	0.029	0.024	0.022	0.02	0.019	yes

Table 5: The answer probability vectors that are used to generate the corresponding answer matrices shown in Table 6.

By using these probabilities, the 25x5-probability matrices for scenarios 2.1, 2.2 and 2.3 (as indicated in Table 3) are then created as follows:

Answer Probability Matrices	
No Contamination	Contamination 1
$A_{25,5} = \begin{bmatrix} 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.20 & 0.50 & 0.20 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.05 & 0.20 & 0.50 & 0.20 \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \end{bmatrix}$	$A_{25,5} = \begin{bmatrix} 0.50 & 0.30 & 0.15 & 0.04 & 0.01 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.50 & 0.30 & 0.15 & 0.04 & 0.01 \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.20 & 0.50 & 0.20 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.05 & 0.20 & 0.50 & 0.20 \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \end{bmatrix}$
Contamination 2	Contamination 3
$A_{25,5} = \begin{bmatrix} 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ 0.50 & 0.30 & 0.15 & 0.04 & 0.01 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.50 & 0.30 & 0.15 & 0.04 & 0.01 \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \end{bmatrix}$	$A_{25,5} = \begin{bmatrix} 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.20 & 0.50 & 0.20 & 0.05 \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.20 & 0.50 & 0.20 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.05 & 0.20 & 0.50 & 0.20 \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & 0.50 & 0.20 & 0.05 & 0.05 \\ 0.50 & 0.30 & 0.15 & 0.04 & 0.01 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.50 & 0.30 & 0.15 & 0.04 & 0.01 \end{bmatrix}$

Table 6: Answer Matrices (5-point scale), with clusters indicated in orange. Each line of vertical dots replaces 3 rows.

As Table 6 shows, the answer probabilities are different for each construct. The difference is not as big as the difference with the contaminated clusters, but it is definitely relevant. This is done to make the simulation setup more realistic. As Arnulf et al. (2018) stated, once the respondent has responded to the an item, the following responses should be inferred by the associations of the items and the framework of the survey, mainly the response categories. In this simulation the structure of the response categories is illustrated by the difference in answer probabilities to each category (construct). Another way of illustrating this is through the correlation matrices.

4.4 Correlation matrices

A correlation matrix is needed to provide the correlation between the answers to the different Likert items. A block correlation matrix is chosen, with each diagonal block consistent and every off-diagonal block, indicating the correlation between the different constructs, is consistent as well. As mentioned, these blocks identify the constructs that each question is linked to. For the base scenario, with only 10 questions, each block consists of two questions to form five constructs. For the scenarios 2 and 3, the correlation matrix has blocks of 5 and 10 questions respectively. So each scenario has 5 constructs, they just differ in size. The Likert items within a construct will have a correlation of 0.7 with each other, and of course 1 with themselves. If the distance between the constructs is 1, meaning the correlation between construct 1 and 2 for example, the correlation will be 0.4. If the distance is 2, 3 or 4, then the correlation will be equal to 0.3, 0.2 or 0.1 respectively. The correlation matrix for Scenario 1 is given below. The matrices for the other scenarios will be given in the Appendix

Correlation Matrices
Scenario 1
$C_{10,10} = \begin{bmatrix} 1.00 & 0.70 & 0.40 & 0.40 & 0.30 & 0.30 & 0.20 & 0.20 & 0.10 & 0.10 \\ 0.70 & 1.00 & 0.40 & 0.40 & 0.30 & 0.30 & 0.20 & 0.20 & 0.10 & 0.10 \\ 0.40 & 0.40 & 1.00 & 0.70 & 0.40 & 0.40 & 0.30 & 0.30 & 0.20 & 0.20 \\ 0.40 & 0.40 & 0.70 & 1.00 & 0.40 & 0.40 & 0.30 & 0.30 & 0.20 & 0.20 \\ 0.30 & 0.30 & 0.40 & 0.40 & 1.00 & 0.70 & 0.40 & 0.40 & 0.30 & 0.30 \\ 0.30 & 0.30 & 0.40 & 0.40 & 0.70 & 1.00 & 0.40 & 0.40 & 0.30 & 0.30 \\ 0.20 & 0.20 & 0.30 & 0.30 & 0.40 & 0.40 & 1.00 & 0.70 & 0.40 & 0.40 \\ 0.20 & 0.20 & 0.30 & 0.30 & 0.40 & 0.40 & 0.70 & 1.00 & 0.40 & 0.40 \\ 0.10 & 0.10 & 0.20 & 0.20 & 0.30 & 0.30 & 0.40 & 0.40 & 1.00 & 0.70 \\ 0.10 & 0.10 & 0.20 & 0.20 & 0.30 & 0.30 & 0.40 & 0.40 & 0.70 & 1.00 \end{bmatrix}$

Table 7: The Correlation Matrix indicating the correlation between the different Likert items and every construct. The within construct correlation is indicated in black, while the between construct correlation is indicated in colors.

Now, everything needed to simulate the scenarios has been outlined. For a specific scenario, for example scenario 2.3, the simulation results will be illustrated according to Figure 1 below. The figures illustrating the other scenarios will be shown in the Appendix.

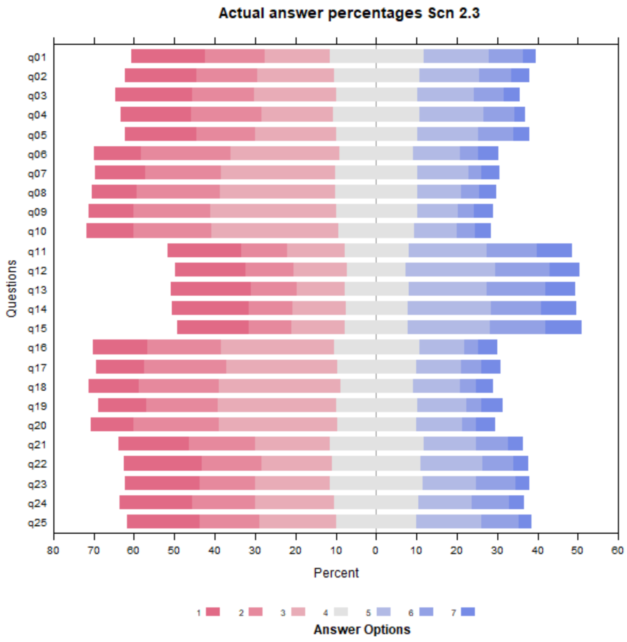
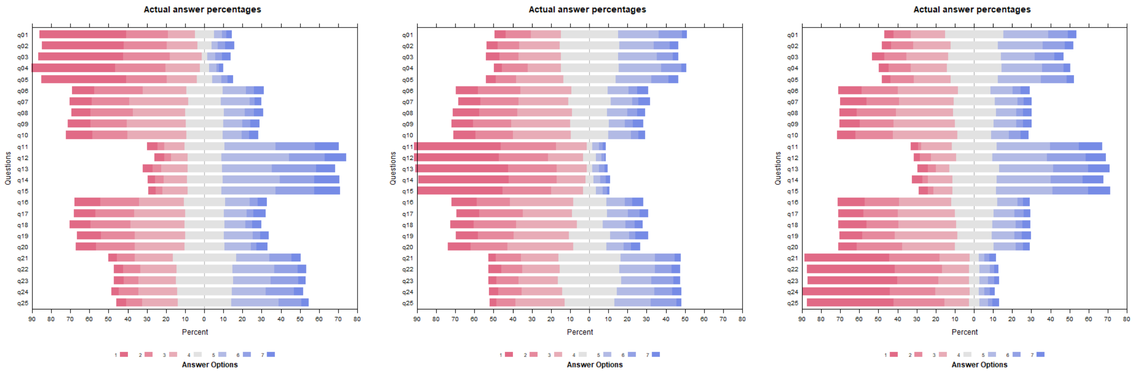


Figure 1: Actual answer percentages for scenario 2.3 with 7-point scale

The clusters at the top, middle and bottom are clearly visible with the answer proportions for those questions skewed in the direction of answer option 1. The question is which observations belong to those clusters. To give an illustration, Figure 2 shows the answer percentages that belong to a group of individuals within each cluster.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

Figure 2: Answer percentages per cluster

Here it is clearly visible that this sub-sample in Cluster 1 is part of the group that forms the cluster that has skewed answers for question 1 to question 5. Similar observations can be made for Cluster 2 and 3. The objective of the ICS method will be to identify those observations and perform clustering in a sufficient way.

4.5 Scatter Pairs

For every scenario introduced in Section 3, there will be a set of six different pairs of scatter matrices considered. These pairs are shown in the table below.

Scatter Pairs		
Pair Number	S1 and S2	Class
pair 1	COV-COV4	1 and 1
pair 2	MCD-COV	3 and 1
pair 3	MCD-COV4	3 and 1
pair 4	TM-COV4	2 and 1
pair 5	TM-COV	2 and 1
pair 6	LCOV-COV	local and 1

Table 8: The different pairs of scatter matrices that will be considered and their corresponding classes of robustness.

In ICS, the use of several combinations of scatter matrices of different classes is often recommended (Tyler et al., 2009)(Nordhausen et al., 2008b), because of the lack of theory on which works best in new scenarios. Archimbaud et al. (2018) suggested the approach of pair 1 and pair 2, however their paper was with the purpose of outlier detection. Therefore, there is no guarantee that they will be able to recover the independent components that are needed. Furthermore, two pairs with the tM estimator are added. As discussed this is a nice alternative to the M-estimator that is presented in the R package, and provides a new estimator from class 2. Finally, the LCOV-COV pair is chosen as it was used in Alfons et al. (2022). In their research, its clustering performance in a continuous data setting was promising and therefore its use is warranted here as well.

4.6 PCA

The term Principal Component Analysis was first introduced by Hotelling (1933). It is a multivariate statistical method that is used in most scientific fields of research (Abdi and Williams, 2010). The method analyzes a data table that contains observations of multiple, correlated, dependent variables. Its goal is to find a new subspace of orthogonal variables called the principal components. This subspace should be of lower dimension but still express the important information from the original data set (Abdi and Williams, 2010).

4.6.1 Setups

For PCA, two setups will be considered. The first setup utilizes the *prcomp*() function from the *stats* package in R (R Core Team, 2020). It is basic PCA with centered data around zero as well as scaled data in order to obtain unit variance before the Principal Component Analysis is actually done.

In the second setup, the *PcaHubert*() function from the *rrcov* package (Todorov and Filzmoser, 2009) in R (R Core Team, 2020) is used in order to apply PCA with a robust estimate of the covariance matrix. The MCD estimator is chosen with an $\alpha = 0.6$. A relatively high α is explained by the fact that issues will arise with the computation of the MCD estimator, especially in lower dimensions. However, since the setup of the simulation is known to lack outliers a higher α can be justified.

4.6.2 Selection of Principal Components

The Principal Components will be selected based on the 80%-criterion as suggested by Alfons et al. (2022). This means that as many components are selected as necessary to explain at least 80% of the variance in the data. It is executed by taking the cumulative sum of the proportion of explained variance and checking for which Principal Component it exceeds 0.8. All the components up to and including that one are selected. Also, the $k-1$ criterion is used, which keeps the first $k-1$ components. Here k is equal to the known number of clusters.

5 Results

In this section the results for the scatter pairs $COV-COV4$, $MCD-COV$, $TM-COV$ and $LCOV-COV$ will be discussed, as well as the relative performance compared to $K-means$ with no dimension reduction and dimension reduction by PCA with $MCD_{0.6}$. This smaller selection of scatter pairs is motivated by the fact that these were the best performing scatters. The scatter pairs that are not presented, $MCD-COV4$ and $TM-COV4$, had very similar performance to $MCD-COV$ and $TM-COV$ respectively. Also, the robust PCA performs significantly better than regular PCA.

5.1 Tandem Clustering with ICS and PCA

In Figure 3 below, the Adjusted Rand Index for every scenario and all scatter pairs are shown in a boxplot. It is clearly visible that the performance of most scatter pairs is not at the level that ICS usually achieves when analyzing continuous data. However, tandem clustering with ICS and $K-means$ is outperforming tandem clustering with PCA as well as applying the $K-means$ algorithm to the original data. This is true for scatter pairs $COV-COV4$ and $TM-COV$, and for all scenarios except Scenario 3.3 with a 5 point Likert scale. PCA is outperforming the ICS methods here, but its performance for all other scenarios and setups seems significantly worse. Also, the component selection for PCA is done by the $k-1$ criterion, which significantly outperforms the 80%-criterion here. The other scatter pairs also outperform PCA and the original data, however, for the scenarios with two clusters, they seem to struggle more. Also, the performance of all scatter pairs seems to be slightly better for the 11-point Likert scale than for the 5-point scale. In terms of standard deviation, the wider Likert scale definitely improves performance, but in terms of height of the ARI it is less clear. However, this improvement is not linear since the 7-point scale seems to present some issues, so it is not clear whether increasing the size of the scale genuinely improves interpretability.

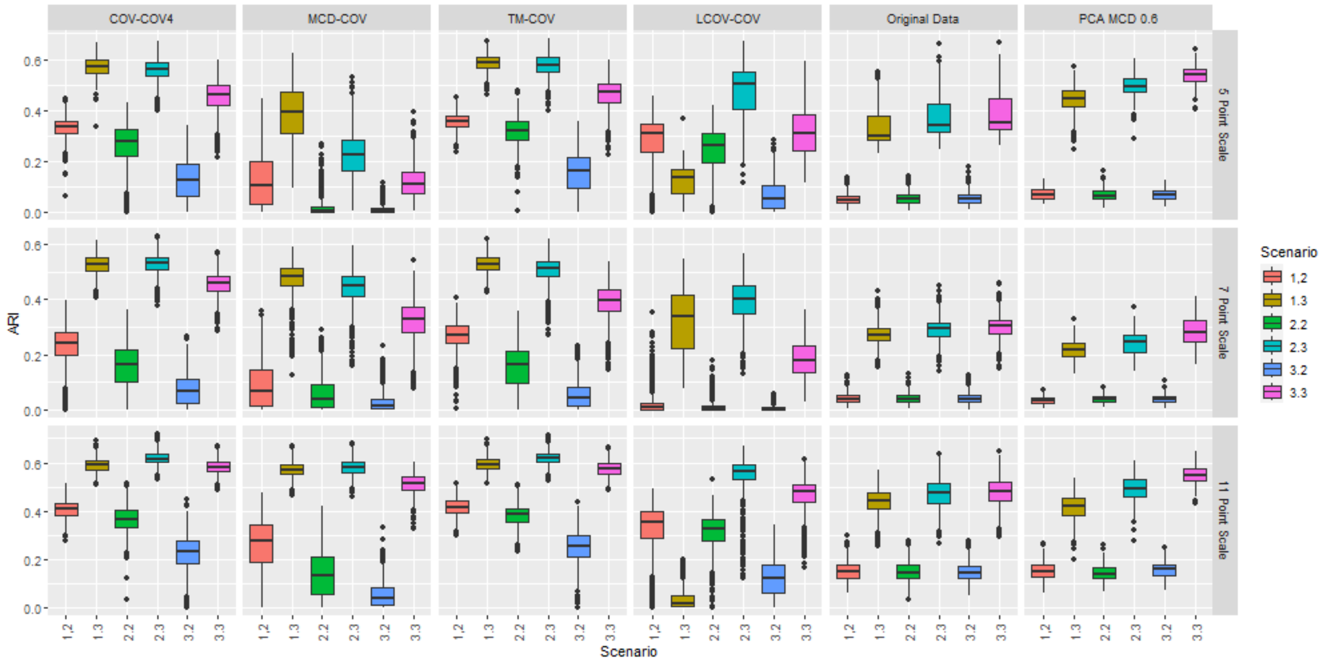


Figure 3: Adjusted Rand Index evaluating performance of all scatter pairs for every scenario. The component selection method for the ICS setups is manual selection. For the PCA setup the $k-1$ criterion is used. This Figure shows the ARIs for the 5-point, 7-point and 11-point Likert Scale.

The figure clearly illustrates that the use of ICS is justified when applying tandem clustering to RSD. Its performance is significantly better than PCA for almost all scenarios. Furthermore, PCA does not seem to significantly improve the ARI in most scenarios, when compared to the original data.

5.2 Component Selection Issues

In order to illustrate the difference between the performance achieved by the different component selection methods, Figure 4 is shown.

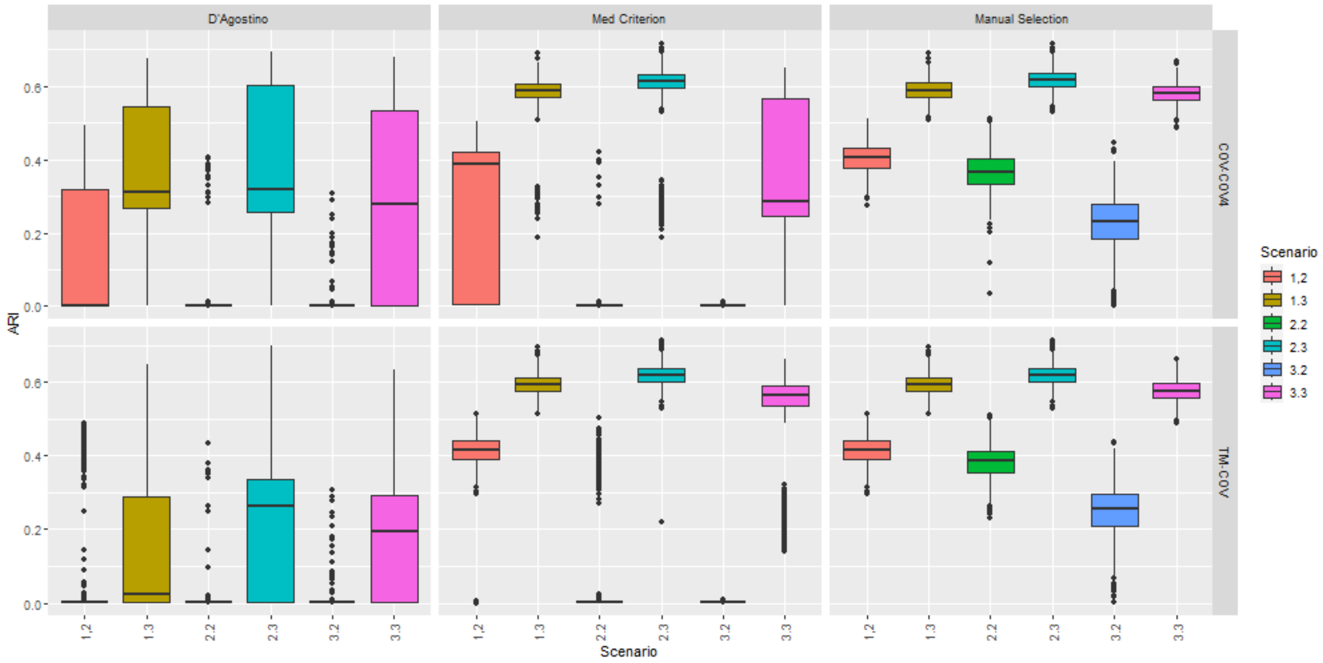


Figure 4: Adjusted Rand Index evaluating performance of the $COV - COV4$ and $TM - COV$ scatter pairs for every scenario. The Figure shows the different performances achieved by the component selection methods for the two best performing ICS scatter pairs. This Figure shows the ARIs for the 11-point Likert Scale.

The performance of the selection of components by the D'Agostino test is relatively bad. Higher ARIs are obtained with both the Med Criterion and Manual Selection. Another thing to note is that the Med Criterion seems to perform relatively well, except for scenarios 2.2 and 3.2, where it clearly selects the wrong components. This issue presents itself for all scatter pairs, however, scenario 1.3 does show decent performance for the Med Criterion. To confirm the bad performance of the D'Agostino test, Table 9 is generated. It shows the mean, over 1000 iterations, of the number of components selected by the D'Agostino test for every scatter pair and for scenarios with 2 or 3 clusters.

Mean of Number of Components Selected by D'Agostino Test							
Scatter Pair	Number of Clusters	Start (5 lvl)	End (5 lvl)	Start (7 lvl)	End (7 lvl)	Start (11 lvl)	End (11 lvl)
COV-COV4	2	2.5	0.3	1.3	0.6	1.8	0.2
COV-COV4	3	2.4	1.0	1.4	0.8	2.0	1.4
MCD-COV	2	2.2	0.8	1.4	0.6	1.6	0.2
MCD-COV	3	2.1	0.3	2.0	0.4	2.4	0.9
TM-COV	2	2.6	0.3	1.2	0.7	1.5	0.2
TM-COV	3	2.5	1.1	1.2	0.7	1.9	1.1
LCOV-COV	2	2.0	0.1	1.3	0.5	1.2	0.2
LCOV-COV	3	2.4	0.6	1.3	0.6	1.8	0.9

Table 9: Mean of the number of components selected from the start and end, by the D'Agostino test. Selecting components from the start or end, means that the D'Agostino test found structure on the first few or last few components.

The D'Agostino test finds the structure in the data mainly on the first few Invariant Coordinates. This is a clear indication of an issue because inspection of the scatter plots of the Invariant Coordinates tells us that the structure is found on the last few coordinates for all scatter pairs. As an example, the scatter plots for the $COV - COV4$ pair are illustrated in Figure 5.

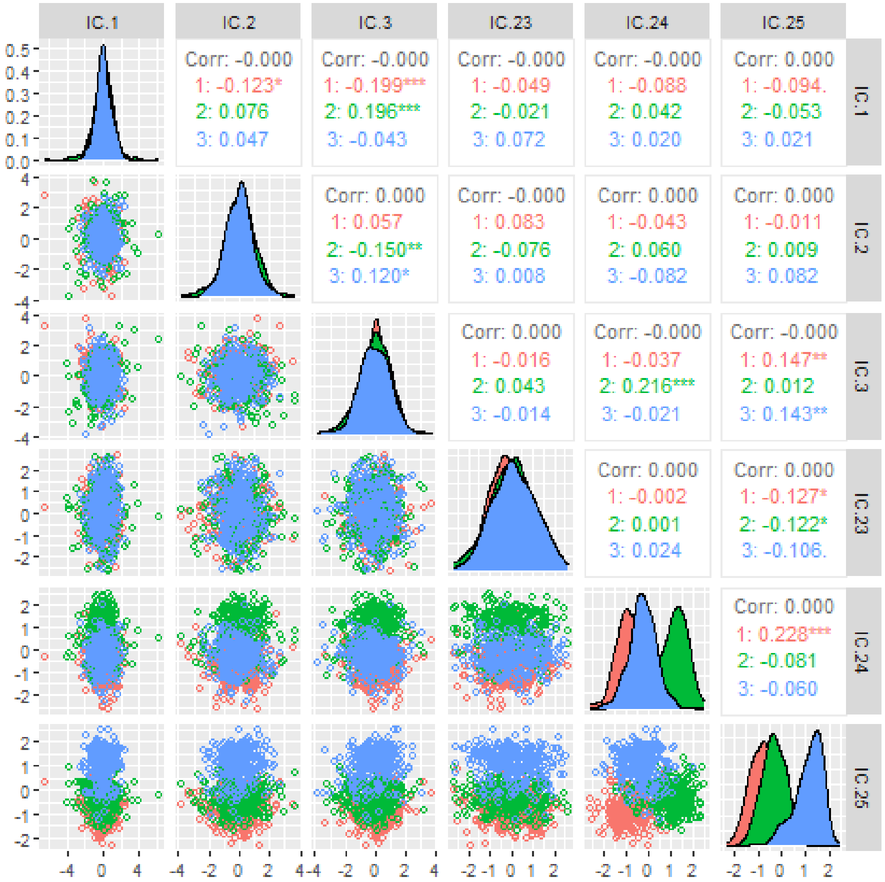
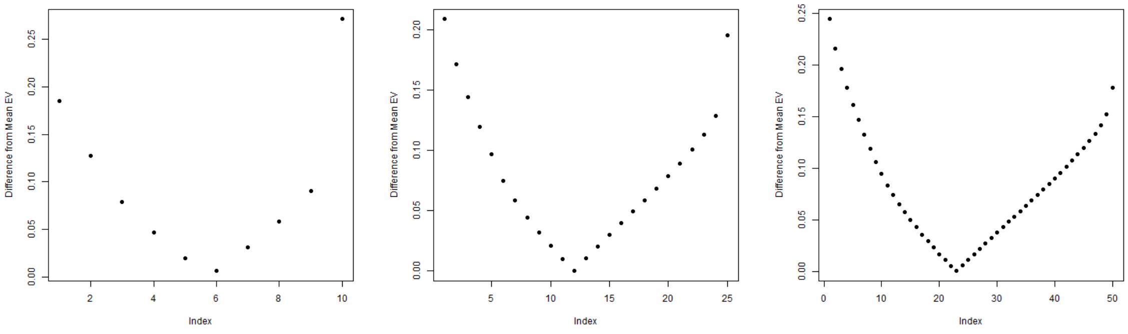


Figure 5: Scatter plots for the Invariant Coordinates, by ICS with the $COV - COV4$ scatter pair. These scatters are set in scenario 2.3 with an 11 point Likert scale.

It is clear that the structure is found on the last 2 coordinates. However, the D'Agostino test, on average, selects the first 2 components as well as the last component, as visible in Table 9. This indicates that the D'Agostino test is not a good selection method for ICS when applied to RSD. Furthermore, the number of coordinates is as expected since this scenario contains 3 clusters. However, the fact that the structure for all scatter pairs is always found on the last few components is unexpected. When performing cluster analysis, Alfons et al.

(2022) used multiple similar scatter pairs and structure was found at the beginning and the end of the ICs. For example, when using the $MCD_{0.5} - COV$ scatter pair for ICS on a real world data set, they find the structure on the first two components.

Another thing to note is that the Med Criterion seems to perform relatively well, except for scenarios 2.2 and 3.2, where it clearly selects the wrong components. This issue presents itself for all scatter pairs but, scenario 1.3 does show decent performance for the Med Criterion. The issues with scenarios 2.2 and 3.2 might have to do with the higher dimensionality of the data, which could be important to note, since the med criterion was succesfully applied in a clustering context with lower dimensions by Alfons et al. (2022). The important difference in setup being that their research contained a data simulation for only 10 variables, whereas this research focuses on 25 and 50 variables as well. Furthermore, they do not investigate whether these criteria work in higher dimensions.



(a) Scenario 1.2

(b) Scenario 2.2

(c) Scenario 3.2

Figure 6: This figure shows the absolute values of the difference between the average eigenvalues of each Invariant Coordinate and their means. The Med Criterion selects the $k-1$ coordinates that are furthest from the mean. The scatter pair is $TM-COV$ and the 11-point Likert scale is used.

Figure 6 shows that, for the scenarios with higher dimensions, the Med Criterion would select the first Invariant Coordinate before the last, since it is further away from the mean. However, it is visible that the last IC comes closer to being selected as the dimension of the data gets lower. This is an indication that the Med Criterion might be better suited as a method when the data is of lower dimension. However, Duembgen et al. (2021) show that refining the estimation of the Invariant Components through a projection pursuit algorithm might solve this issue as well.

5.3 Averaging Constructs

Since the achieved ARI is lower than in the existing literature on ICS applied on continuous data, one idea to improve the performance could be to make the data more continuous. The suggestion by Harpe (2015) to average the data over its constructs is an interesting solution.

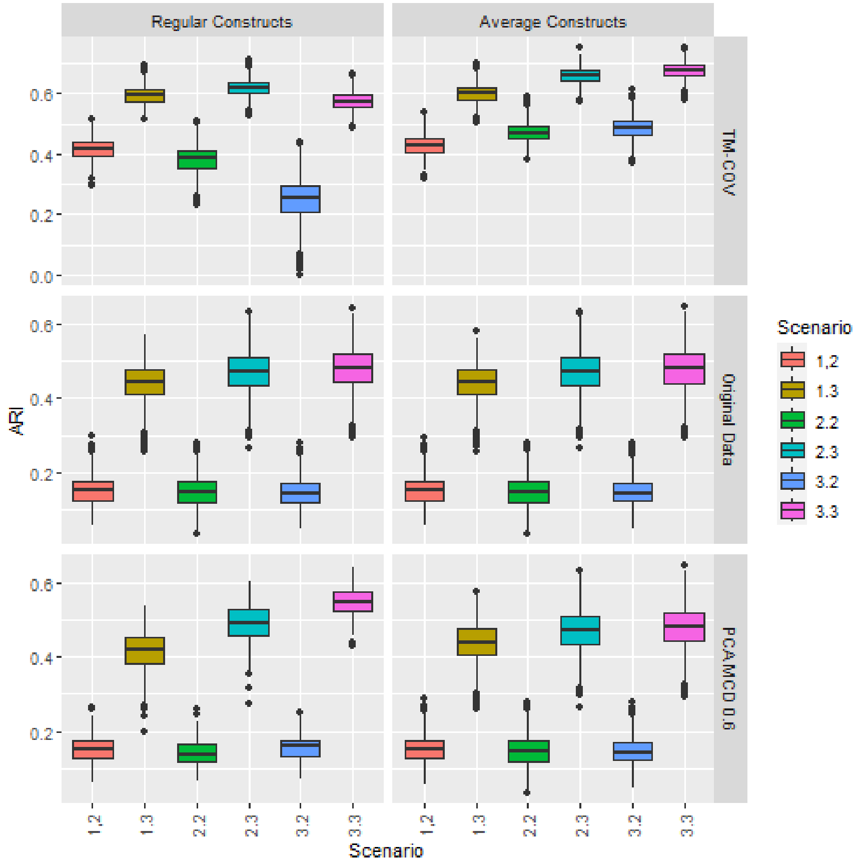


Figure 7: Adjusted Rand Index, difference between the performance of the $TM - COV$ scatter pair, the original data, and tandem clustering with PCA, for each scenario, in the context of averaging clusters and not averaging clusters and an 11 point Likert scale. Coordinates for ICS are selected manually, while components for PCA are selected by the $k - 1$ criterion for the regular constructs and by the 80%-criterion for the average constructs.

The comparison done in Figure 7 shows that for the best performing scatter pair, $TM-COV$, averaging the constructs helps to improve the performance of tandem clustering with ICS. The relative performance of the $K - means$ on the original data and the PCA method is shown in the context of regular and average constructs. Notable is the lack of improvement for the original data and the decrease in performance for PCA, while ICS improves remarkably. The selection of the Principal Components is done by the criterion that obtains optimal results for every setup. For the average constructs it is the 80%-criterion while for the regular constructs it remains the $k-1$ criterion. Averaging the constructs seems to complicates the process of finding the relevant structure for PCA, while it simplifies the process for ICS.

The achieved ARI by applying $K-means$ to the original data is still significantly lower than the ARI for every ICS scatter in every scenario and for all Likert scale levels, when the values for the original data are compared to the values for the different ICS scatter pairs. These values are illustrated in Table 15 in the Appendix. Furthermore, it is of particular importance that ICS is outperforming the robust PCA method for all Likert levels and almost all scatter pair combinations, especially in the setup where constructs are averaged. Another remark that has to be made is that the $K-means$ for the 11 point Likert scale is performing the best in terms of height of achieved ARI and also in terms of standard deviation.

5.4 Discussion and Conclusion

The application of ICS to RSD proves to be an interesting topic of further research. The initial results were not impressive when comparing them to the existing literature on ICS and its clustering purposes as well as the literature on clustering of RSD in general. However, the averaging of constructs seems to be a very important step in the process of making the RSD more analyzable through ICS and dimension reduction in general. It shows that tandem clustering through ICS is definitely an option in this context. The performance measured by the ARI improved significantly for all analyzed scatter pairs, as well as for the selection methods. Furthermore, ICS achieved a higher ARI than PCA and the original data for all scatter pairs. The achieved ARI in many scenarios was higher than in existing research. For example, Jacques and Biernacki (2018) had a similar setup of simulation with mixed and unmixed clusters. Their Adjusted Rand Index achieved with a co-clustering method was 0.99 for the unmixed clusters and 0.58 for the more mixed clusters. ICS for certain scatter pairs is able to achieve close to 0.7 for mixed clusters (scenarios 1.3, 2.3 and 3.3) and close to 0.5 for even more mixed clusters (scenarios 1.2, 2.2 and 3.2). These values cannot be compared directly because of differences in simulation, however, the decrease in

ARI between the scenarios with mixed and unmixed clusters can be compared. In the co-clustering method, that decrease was significantly higher than in tandem clustering with ICS.

The D'Agostino test seems unable to make the correct selections of the ICs, especially if the constructs are not averaged. The Med criterion performs better with component selection and its previous application in this field as well as the indication that it could work more reliably in lower dimensions proves that it demands further research. An applicable and automized selection method in the context of applying ICS on RSD is a relevant step that needs to be made. Moreover, further research into the structure of the Invariant Coordinates is warranted. In this thesis, structure was only found on the last few coordinates for all analyzed scatter pairs[‡], which is in contrast with previous research into ICS and its clustering purposes.

The scatter pairs $COV - COV4$ and $TM - COV$ are performing relatively well. They consistently outperform $LCOV - COV$ and $MCD - COV$ if the correct components are selected (manual selection). Choosing relevant and correct scatter pairs remains an important and current topic of research. A recommendation for further research into applying ICS on RSD is to investigate a wider variety of scatter pairs and to focus on wider Likert scales (11-point), because this generally seems to decrease the standard deviation of the ARI meaning the method performs more consistently. Also, finding a consistently performing Coordinate Selection method is desired. The D'Agostino test and Med criterion analyzed in this research are unable to select the relevant Invariant Coordinates in most scenarios. A selection method that is robust to different setups such as a variety of chosen scatter pairs is needed in order to disregard the time consuming process of manual selection of the coordinates. This could prove to be the key to enhance the performance of ICS in its application to RSD.

[‡] The only exception is the case where constructs are averaged and the chosen scatter pair is $LCOV - COV$. However, this seems not very relevant since it was not performing very well in that scenario anyway.

REFERENCES

H. Abdi and L. J. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

A. Agresti. *Categorical data analysis.* John Wiley & Sons, 2003.

A. Alfons, A. Archimbaud, K. Nordhausen, and A. Ruiz-Gazen. Tandem clustering with invariant coordinate selection. *arXiv preprint arXiv:2212.06108*, 2022.

P. Arabie. Cluster analysis in marketing research. *Advanced methods of marketing research*, pages 160–189, 1994.

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

J. K. Arnulf, K. R. Larsen, and Ø. L. Martinsen. Respondent robotics: simulating responses to likert-scale survey items. *Sage Open*, 8(1):2158244018764803, 2018.

C. A. Bernaards and K. Sijtsma. Bias of factor loadings from questionnaire data with imputed scores. *Journal of Statistical Computation and Simulation*, 75(1):13–23, 2005.

N. Blaikie. *Analyzing quantitative data: From description to explanation.* Sage, 2003.

T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101. URL https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

J. Carifio and R. Perla. Resolving the 50-year debate around using and misusing likert scales, 2008.

J. Carifio and R. J. Perla. Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of social sciences*, 3(3):106–116, 2007.

E. A. Cator and H. P. Lopuhaä. Central limit theorem and influence function for the mcd estimators at general multivariate distributions. *Bernoulli*, 18(2):520–551, 2012.

H. Caussinus and A. Ruiz-Gazen. Interesting projections of multidimensional data by means of generalized principal component analyses. In *Compstat*, pages 121–126. Springer, 1990.

H. Caussinus, M. Fekri, S. Hakam, and A. Ruiz-Gazen. A monitoring display of multivariate outliers. *Computational Statistics & Data Analysis*, 44(1-2):237–252, 2003.

H. Cramér. *Mathematical methods of statistics*, volume 26. Princeton university press, 1999.

D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

P. L. Davies. Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, pages 1269–1292, 1987.

I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98, 2003.

L. Duembgen, K. Gysel, and F. Perler. Refining invariant coordinate selection via local projection pursuit. *arXiv preprint arXiv:2112.11998*, 2021.

L. Dümbgen and D. E. Tyler. On the breakdown properties of some multivariate m-functionals. *Scandinavian Journal of Statistics*, 32(2):247–264, 2005.

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

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2): 179–188, 1936.

K. Goldfeld and J. Wujciak-Jens. simstudy: Illuminating research methods through data generation. *Journal of Open Source Software*, 5(54):2763, 2020.

S. E. Harpe. How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6):836–850, 2015.

L. L. Havlicek and N. L. Peterson. Robustness of the pearson correlation against violations of assumptions. *Perceptual and Motor Skills*, 43(3_suppl):1319–1334, 1976.

C. Hennig. Discussion of" invariant coordinate selection", by de tyler, f. critchley, l. dümbgen, and h. oja. *Journal of the Royal Statistical Society. Series B. Methodological*, 71:579–583, 2009.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

P. J. Huber. Robust statistics. New York: Wiley, 1981.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

J. Jacques and C. Biernacki. Model-based co-clustering for ordinal data. *Computational Statistics & Data Analysis*, 123:101–115, 2018.

S. Jamieson. Likert scales: How to (ab) use them? *Medical education*, 38(12):1217–1218, 2004.

J. T. Kent, D. E. Tyler, and Y. Vard. A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics-Simulation and Computation*, 23(2):441–453, 1994.

T. R. Knapp. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research*, 39(2):121–123, 1990.

W. Kuzon, M. Urbanchek, and S. McCabe. The seven deadly sins of statistical analysis. *Annals of plastic surgery*, 37:265–272, 1996.

S. Labovitz. The assignment of numbers to rank order categories. *American sociological review*, pages 515–524, 1970.

S.-O. Leung. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of social service research*, 37(4):412–421, 2011.

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

R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

H. P. Lopuhaa. On the relation between s-estimators and m-estimators of multivariate location and covariance. *The Annals of Statistics*, pages 1662–1683, 1989.

R. A. Maronna. Robust m-estimators of multivariate location and scatter. *The annals of statistics*, pages 51–67, 1976.

T. H. Matta, L. Rutkowski, D. Rutkowski, and Y.-L. Liaw. lsasim: an r package for simulating large-scale assessment data. *Large-scale Assessments in Education*, 6(1):1–33, 2018.

T. Nemoto and D. Beglar. Likert-scale questionnaires. In *JALT 2013 conference proceedings*, pages 1–8, 2014.

K. Nordhausen, H. Oja, and E. Ollila. Robust independent component analysis based on two scatter matrices. *Austrian Journal of Statistics*, 37(1):91–100, 2008a.

K. Nordhausen, H. Oja, and D. E. Tyler. Tools for exploring multivariate data: The package ICS. *Journal of Statistical Software*, 28(6):1–31, 2008b. URL `http://www.jstatsoft.org/v28/i06/`.

G. Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

E. S. Pearson. The analysis of variance in cases of non-normal variation. *Biometrika*, pages 114–133, 1931.

E. S. Pearson. The test of significance for the correlation coefficient. *Journal of the American Statistical Association*, 27(174):128–134, 1932a.

E. S. Pearson. The test of significance for the correlation coefficient: Some further results. *Journal of the American Statistical Association*, 27(180):424–426, 1932b.

G. Pell. Use and misuse of likert scales. 2005.

M. A. Pett. *Nonparametric statistics for health care research: Statistics for small samples and unusual distributions.* Sage Publications, 2015.

J. Podani. Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions. *Journal of Vegetation Science*, 16(5):497–510, 2005.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

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

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

S. S. Stevens. Mathematics, measurement, and psychophysics. 1951.

G. M. Sullivan and A. R. Artino Jr. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542, 2013.

V. Todorov and P. Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009. doi: 10.18637/jss.v032.i03.

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

M. Van de Velden, A. I. D'Enza, and F. Palumbo. Cluster correspondence analysis. *Psychometrika*, 82:158–185, 2017.

M. Walesiak and A. Dudek. Finding groups in ordinal data: an examination of some clustering procedures. In *Classification as a Tool for Research: Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation eV, Dresden, March 13-18, 2009*, pages 185–192. Springer, 2010.

H. Wu and S.-O. Leung. Can likert scales be treated as interval scales?—a simulation study. *Journal of Social Service Research*, 43(4):527–532, 2017.

Appendix

Equations

To illustrate the MPLT model, we assume there are N respondents, each indexed by i, that answered to J Likert items, each indexed by j, with answer categories indicated by $x = (0, \ldots, r)$. They MPLT model is then of the form

$$P(X_{ij} = x|\theta_{i1}, \ldots, \theta_{iQ}) = \frac{\exp\left(\sum_{q=1}^{Q} \theta_{iq} B_{jq} x - \Psi_{jx}\right)}{\sum_{y=0}^{r} \exp\left(\sum_{q=1}^{Q} \theta_{iq} B_{jq} y - \Psi_{jy}\right)}, \tag{9}$$

where there are Q latent traits that are indexed by q and denoted as $\Theta = (\theta_1, \ldots, \theta_Q)$. Attached to each latent trait is a wait B for every item j given by $\mathbf{B}_j = (B_{j1}, \ldots, B_{jQ})$ where B_{jq} indicates the impact of latent trait q on item j. Lastly, Ψ_{jx} is a separation parameter for item j on answer category x. Equation 9 thus indicates the probability of an individual i responding to Likert item j with answer x, indicated by $P(X_{ij} = x|\theta_{i1}, \ldots, \theta_{iQ})$. This type of model is a logit-link model (Agresti, 2003) and as stated by Bernaards and Sijtsma (2005, p. 15) "such models are called divide-by-total models or adjacent category models".

Tables

This Table shows the correlation matrices for the scenarios with 25 and 50 dimensions. The correlation within each construct is shown in a specific color. The correlation between constructs decreases if the constructs are situated further apart.

Correlation Matrices
Scenario 2.1-2.3
$C_{25,25} = \begin{bmatrix} 1.00 & \cdots & 0.70 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 & 0.20 & \cdots & 0.20 & 0.10 & \cdots & 0.10 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.70 & \cdots & 1.00 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 & 0.20 & \cdots & 0.20 & 0.10 & \cdots & 0.10 \\ 0.40 & \cdots & 0.40 & 1.00 & \cdots & 0.70 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 & 0.20 & \cdots & 0.20 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.40 & \cdots & 0.40 & 0.70 & \cdots & 1.00 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 & 0.20 & \cdots & 0.20 \\ 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 1.00 & \cdots & 0.70 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 0.70 & \cdots & 1.00 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 \\ 0.20 & \cdots & 0.20 & 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 1.00 & \cdots & 0.70 & 0.40 & \cdots & 0.40 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & \cdots & 0.20 & 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 0.70 & \cdots & 1.00 & 0.40 & \cdots & 0.40 \\ 0.10 & \cdots & 0.10 & 0.20 & \cdots & 0.20 & 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 1.00 & \cdots & 0.70 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.10 & \cdots & 0.10 & 0.20 & \cdots & 0.20 & 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 0.70 & \cdots & 1.00 \end{bmatrix}$
Scenario 3.1-3.3
$C_{50,50} = \begin{bmatrix} 1.00 & \cdots & 0.70 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 & 0.20 & \cdots & 0.20 & 0.10 & \cdots & 0.10 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.70 & \cdots & 1.00 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 & 0.20 & \cdots & 0.20 & 0.10 & \cdots & 0.10 \\ 0.40 & \cdots & 0.40 & 1.00 & \cdots & 0.70 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 & 0.20 & \cdots & 0.20 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.40 & \cdots & 0.40 & 0.70 & \cdots & 1.00 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 & 0.20 & \cdots & 0.20 \\ 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 1.00 & \cdots & 0.70 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 0.70 & \cdots & 1.00 & 0.40 & \cdots & 0.40 & 0.30 & \cdots & 0.30 \\ 0.20 & \cdots & 0.20 & 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 1.00 & \cdots & 0.70 & 0.40 & \cdots & 0.40 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.20 & \cdots & 0.20 & 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 0.70 & \cdots & 1.00 & 0.40 & \cdots & 0.40 \\ 0.10 & \cdots & 0.10 & 0.20 & \cdots & 0.20 & 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 1.00 & \cdots & 0.70 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.10 & \cdots & 0.10 & 0.20 & \cdots & 0.20 & 0.30 & \cdots & 0.30 & 0.40 & \cdots & 0.40 & 0.70 & \cdots & 1.00 \end{bmatrix}$

Table 10: Correlation Matrices, every line of vertical dots in $C_{25,25}$ replaces 3 rows in the table, every line of vertical dots in $C_{50,50}$ replaces 8 rows in the table.

The following tables show the mean ARI for the scenarios 2.2-3.3, for all three selection methods.

Mean ARI for each Scenario (D'Agostino Test)				
Scatter Pair	Scenario	5 lvl	7 lvl	11 lvl
Original Data	2.2	0.05	0.04	0.14
Original Data	2.3	0.37	0.29	0.47
Original Data	3.2	0.05	0.04	0.14
Original Data	3.3	0.38	0.30	0.47
COV-COV4	2.2	0.02	0.05	0.00
COV-COV4	2.3	0.32	0.19	0.35
COV-COV4	3.2	0.00	0.00	0.00
COV-COV4	3.3	0.24	0.12	0.28
MCD-COV	2.2	0.00	0.00	0.00
MCD-COV	2.3	0.02	0.05	0.08
MCD-COV	3.2	0.00	0.00	0.00
MCD-COV	3.3	0.00	0.02	0.04
TM-COV	2.2	0.00	0.03	0.00
TM-COV	2.3	0.17	0.18	0.23
TM-COV	3.2	0.00	0.01	0.00
TM-COV	3.3	0.17	0.09	0.19
LCOV-COV	2.2	0.00	0.00	0.00
LCOV-COV	2.3	0.03	0.10	0.05
LCOV-COV	3.2	0.00	0.00	0.00
LCOV-COV	3.3	0.01	0.02	0.04
ROBPCA	2.2	0.05	0.04	0.14
ROBPCA	2.3	0.37	0.29	0.47
ROBPCA	3.2	0.05	0.04	0.14
ROBPCA	3.3	0.38	0.30	0.47

Table 11: Mean of the ARI over 1000 iterations for each scenario. Selection of the Invariant Components for the ICS scenarios is done by the D'Agostino Test

Mean ARI for each Scenario (Med Criterion)				
Scatter Pair	Scenario	5 lvl	7 lvl	11 lvl
COV-COV4	2.2	0.04	0.00	0.00
COV-COV4	2.3	0.33	0.43	0.59
COV-COV4	3.2	0.00	0.00	0.00
COV-COV4	3.3	0.16	0.17	0.36
MCD-COV	2.2	0.00	0.00	0.00
MCD-COV	2.3	0.00	0.00	0.00
MCD-COV	3.2	0.00	0.00	0.00
MCD-COV	3.3	0.00	0.00	0.00
TM-COV	2.2	0.15	0.00	0.04
TM-COV	2.3	0.40	0.48	0.61
TM-COV	3.2	0.01	0.00	0.00
TM-COV	3.3	0.24	0.24	0.50
LCOV-COV	2.2	0.00	0.00	0.00
LCOV-COV	2.3	0.00	0.01	0.01
LCOV-COV	3.2	0.00	0.00	0.00
LCOV-COV	3.3	0.00	0.00	0.00

Table 12: Mean of the ARI over 1000 iterations for each scenario. Selection of the Invariant Components for the ICS scenarios is done by the Med Criterion.

Mean ARI for each Scenario (Manual Selection)				
Scatter Pair	Scenario	5 lvl	7 lvl	11 lvl
COV-COV4	2.2	0.26	0.15	0.36
COV-COV4	2.3	0.55	0.52	0.61
COV-COV4	3.2	0.12	0.07	0.22
COV-COV4	3.3	0.45	0.45	0.58
MCD-COV	2.2	0.02	0.05	0.13
MCD-COV	2.3	0.22	0.44	0.58
MCD-COV	3.2	0.00	0.02	0.05
MCD-COV	3.3	0.11	0.31	0.51
TM-COV	2.2	0.31	0.15	0.38
TM-COV	2.3	0.57	0.50	0.61
TM-COV	3.2	0.15	0.05	0.24
TM-COV	3.3	0.46	0.39	0.57
LCOV-COV	2.2	0.24	0.01	0.31
LCOV-COV	2.3	0.46	0.39	0.54
LCOV-COV	3.2	0.06	0.00	0.12
LCOV-COV	3.3	0.31	0.18	0.46

Table 13: Mean of the ARI over 1000 iterations for each scenario. Selection of the Invariant Components for the ICS scenarios is done by Manual Selection. This means that for each scenario $k-1$ components are selected. This is done from the end of the ICS coordinates because all structure is visible there, as seen in the scatter plots.

The three following tables show the mean ARI for the three selection methods in the context of average constructs.

Mean ARI for each Scenario (D'Agostino Test & Average Constructs)				
Scatter Pair	Scenario	5 lvl	7 lvl	11 lvl
Original Data	1.2	0.19	0.04	0.14
Original Data	1.3	0.33	0.29	0.47
Original Data	2.2	0.05	0.04	0.14
Original Data	2.3	0.37	0.29	0.47
Original Data	3.2	0.05	0.04	0.14
Original Data	3.3	0.38	0.29	0.47
COV-COV4	1.2	0.19	0.27	0.15
COV-COV4	1.3	0.34	0.24	0.33
COV-COV4	2.2	0.12	0.29	0.03
COV-COV4	2.3	0.40	0.31	0.38
COV-COV4	3.2	0.10	0.30	0.01
COV-COV4	3.3	0.42	0.35	0.40
MCD-COV	1.2	0.00	0.03	0.03
MCD-COV	1.3	0.04	0.11	0.17
MCD-COV	2.2	0.00	0.03	0.01
MCD-COV	2.3	0.03	0.13	0.20
MCD-COV	3.2	0.00	0.03	0.01
MCD-COV	3.3	0.04	0.16	0.25
TM-COV	1.2	0.02	0.08	0.07
TM-COV	1.3	0.04	0.09	0.10
TM-COV	2.2	0.01	0.08	0.01
TM-COV	2.3	0.04	0.14	0.16
TM-COV	3.2	0.01	0.09	0.00
TM-COV	3.3	0.05	0.18	0.18
LCOV-COV	1.2	0.02	0.20	0.10
LCOV-COV	1.3	0.37	0.36	0.37
LCOV-COV	2.2	0.02	0.22	0.14
LCOV-COV	2.3	0.47	0.40	0.40
LCOV-COV	3.2	0.03	0.23	0.16
LCOV-COV	3.3	0.49	0.42	0.42
ROBPCA	1.2	0.05	0.04	0.14
ROBPCA	1.3	0.37	0.29	0.47
ROBPCA	2.2	0.05	0.04	0.14
ROBPCA	2.3	0.37	0.29	0.47
ROBPCA	3.2	0.05	0.04	0.14
ROBPCA	3.3	0.38	0.30	0.47

Table 14: Mean of the ARI over 1000 iterations for each scenario. The constructs of the data are averaged. Selection of the Invariant Components for the ICS scenarios is done by the D'Agostino Test. For the original data no component selection is needed and for the robust PCA component selection is done by the 80%-criterion.

Mean ARI for each Scenario (Manual Selection & Average Constructs)				
Scatter Pair	Scenario	5 lvl	7 lvl	11 lvl
Original Data	1.2	0.19	0.04	0.14
Original Data	1.3	0.33	0.29	0.47
Original Data	2.2	0.05	0.04	0.14
Original Data	2.3	0.37	0.29	0.47
Original Data	3.2	0.05	0.04	0.14
Original Data	3.3	0.38	0.29	0.47
COV-COV4	2.2	0.36	0.28	0.42
COV-COV4	2.3	0.59	0.53	0.59
COV-COV4	2.2	0.41	0.31	0.46
COV-COV4	2.3	0.66	0.59	0.65
COV-COV4	3.2	0.44	0.33	0.47
COV-COV4	3.3	0.69	0.62	0.67
MCD-COV	2.2	0.28	0.11	0.38
MCD-COV	2.3	0.51	0.51	0.58
MCD-COV	2.2	0.36	0.15	0.43
MCD-COV	2.3	0.58	0.57	0.64
MCD-COV	3.2	0.38	0.16	0.44
MCD-COV	3.3	0.62	0.60	0.66
TM-COV	2.2	0.37	0.30	0.43
TM-COV	2.3	0.60	0.54	0.60
TM-COV	2.2	0.42	0.34	0.46
TM-COV	2.3	0.67	0.60	0.65
TM-COV	3.2	0.45	0.36	0.48
TM-COV	3.3	0.70	0.63	0.67
LCOV-COV	2.2	0.00	0.16	0.08
LCOV-COV	2.3	0.35	0.41	0.56
LCOV-COV	2.2	0.01	0.17	0.12
LCOV-COV	2.3	0.47	0.53	0.63
LCOV-COV	3.2	0.01	0.19	0.16
LCOV-COV	3.3	0.55	0.59	0.66
ROBPCA	1.2	0.05	0.04	0.14
ROBPCA	1.3	0.37	0.29	0.47
ROBPCA	2.2	0.05	0.04	0.14
ROBPCA	2.3	0.37	0.29	0.47
ROBPCA	3.2	0.05	0.04	0.14
ROBPCA	3.3	0.38	0.30	0.47

Table 15: Mean of the ARI over 1000 iterations for each scenario. The constructs of the data are averaged. Selection of the Invariant Components for the ICS scenarios is done by Manual Selection. In the setup where the constructs are averaged, structure for all scatter pairs is still found at the end of the Invariant Components, except for the LCOV-COV scatter pair. In the averaging construct setup, and for that scatter pair, the structure is found at the start of the Invariant Components.

Mean ARI for each Scenario (Med Criterion & Average Constructs)				
Scatter Pair	Scenario	5 lvl	7 lvl	11 lvl
Original Data	1.2	0.19	0.04	0.14
Original Data	1.3	0.33	0.29	0.47
Original Data	2.2	0.05	0.04	0.14
Original Data	2.3	0.37	0.29	0.47
Original Data	3.2	0.05	0.04	0.14
Original Data	3.3	0.38	0.29	0.47
COV-COV4	2.2	0.31	0.20	0.42
COV-COV4	2.3	0.13	0.20	0.44
COV-COV4	2.2	0.36	0.24	0.45
COV-COV4	2.3	0.16	0.23	0.50
COV-COV4	3.2	0.39	0.26	0.47
COV-COV4	3.3	0.17	0.25	0.53
MCD-COV	2.2	0.09	0.01	0.09
MCD-COV	2.3	0.00	0.13	0.12
MCD-COV	2.2	0.09	0.01	0.08
MCD-COV	2.3	0.00	0.14	0.14
MCD-COV	3.2	0.08	0.02	0.08
MCD-COV	3.3	0.00	0.16	0.18
TM-COV	2.2	0.36	0.28	0.43
TM-COV	2.3	0.09	0.25	0.39
TM-COV	2.2	0.40	0.32	0.46
TM-COV	2.3	0.14	0.28	0.48
TM-COV	3.2	0.43	0.34	0.48
TM-COV	3.3	0.16	0.30	0.52
LCOV-COV	2.2	0.04	0.13	0.07
LCOV-COV	2.3	0.34	0.35	0.39
LCOV-COV	2.2	0.02	0.14	0.11
LCOV-COV	2.3	0.39	0.43	0.53
LCOV-COV	3.2	0.01	0.16	0.14
LCOV-COV	3.3	0.42	0.49	0.59
ROBPCA	1.2	0.05	0.04	0.14
ROBPCA	1.3	0.37	0.29	0.47
ROBPCA	2.2	0.05	0.04	0.14
ROBPCA	2.3	0.37	0.29	0.47
ROBPCA	3.2	0.05	0.04	0.14
ROBPCA	3.3	0.38	0.30	0.47

Table 16: Mean of the ARI over 1000 iterations for each scenario. The constructs of the data are averaged. Selection of the Invariant Components for the ICS scenarios is done by the Med Criterion.

Figures

The following figures show the actual answer percentages over the different scenarios indicated in the captions.

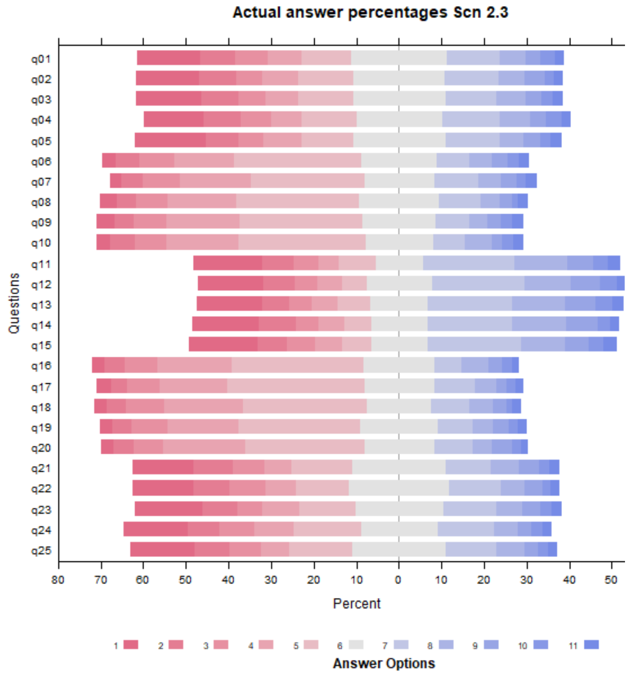


Figure 8: Scenario 2.3 with 11-point scale

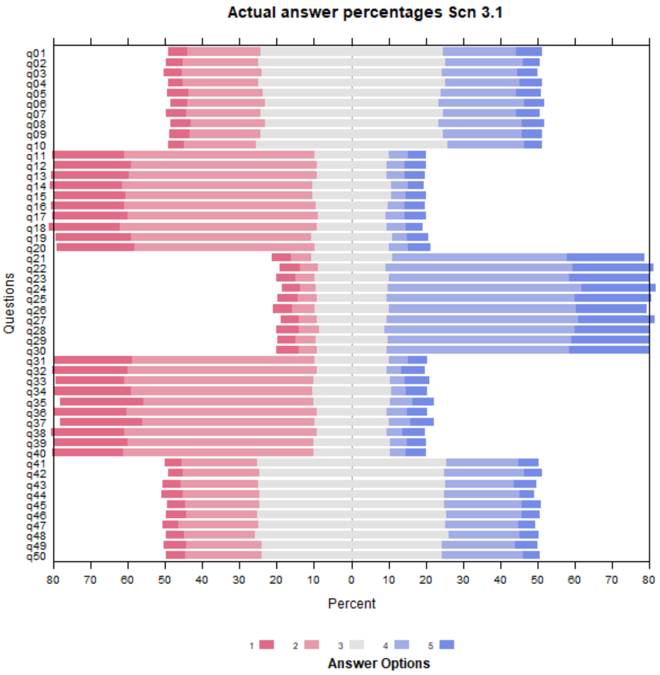


Figure 9: Scenario 3.1 with 5-point scale

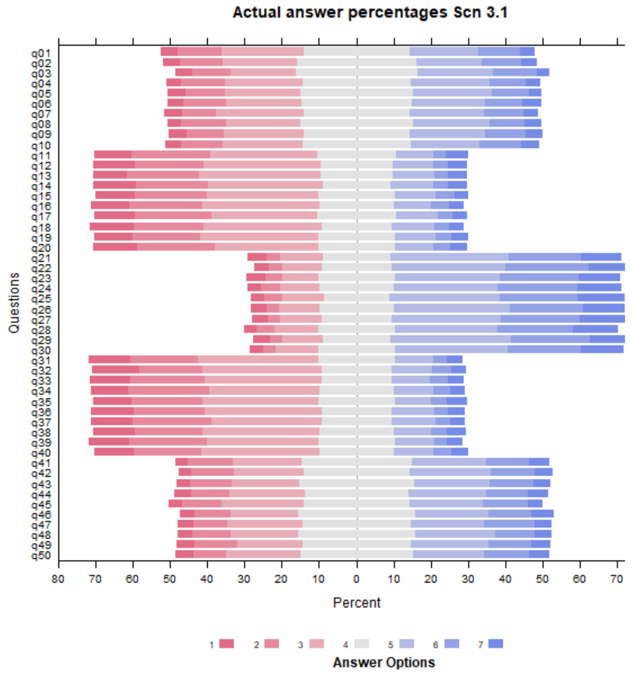


Figure 10: Scenario 3.1 with 7-point scale

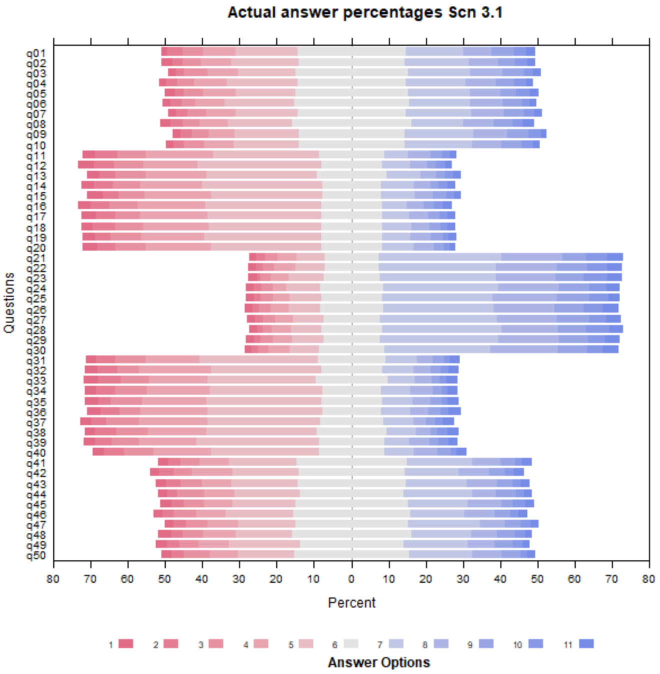


Figure 11: Scenario 3.1 with 11-point scale

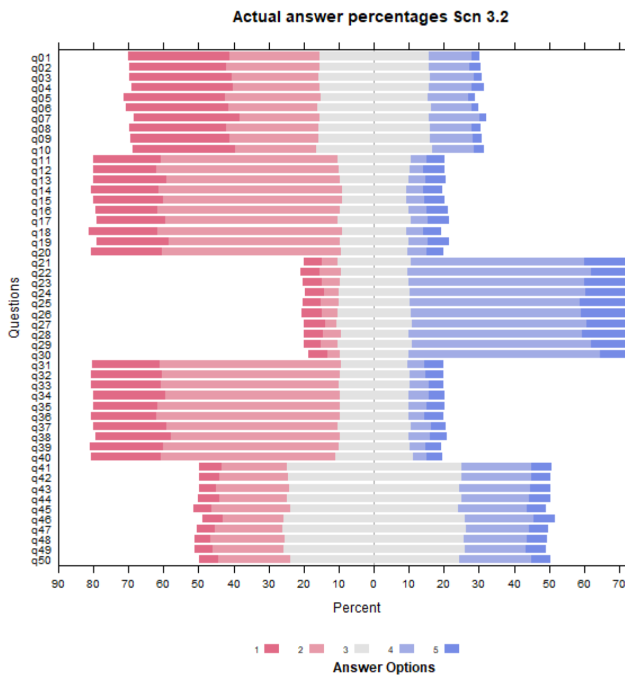


Figure 12: Scenario 3.2 with 5-point scale

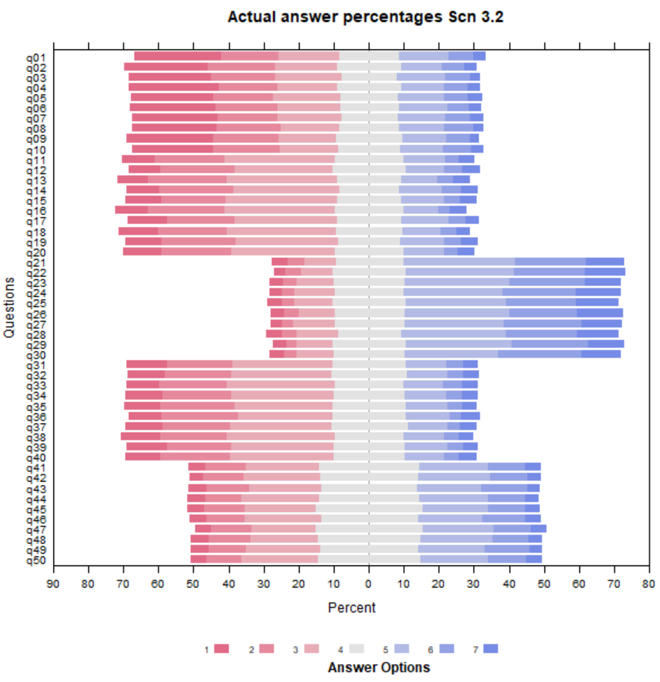


Figure 13: Scenario 3.2 with 7-point scale

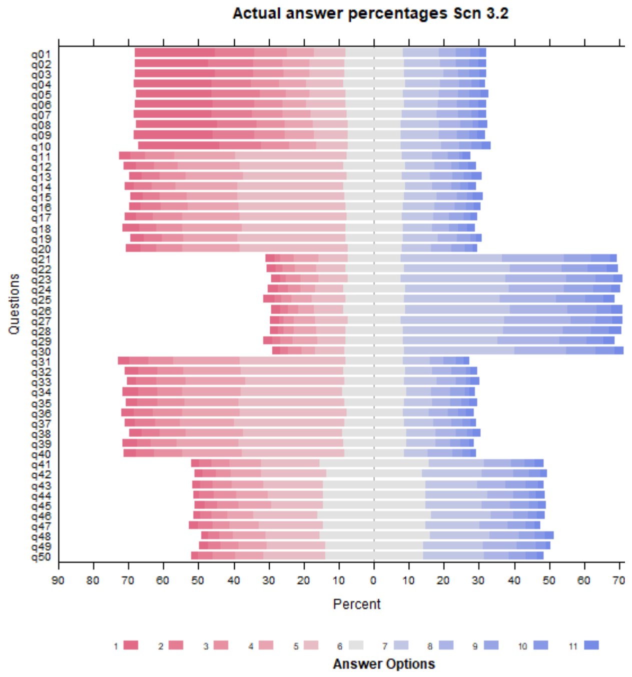


Figure 14: Scenario 3.2 with 11-point scale

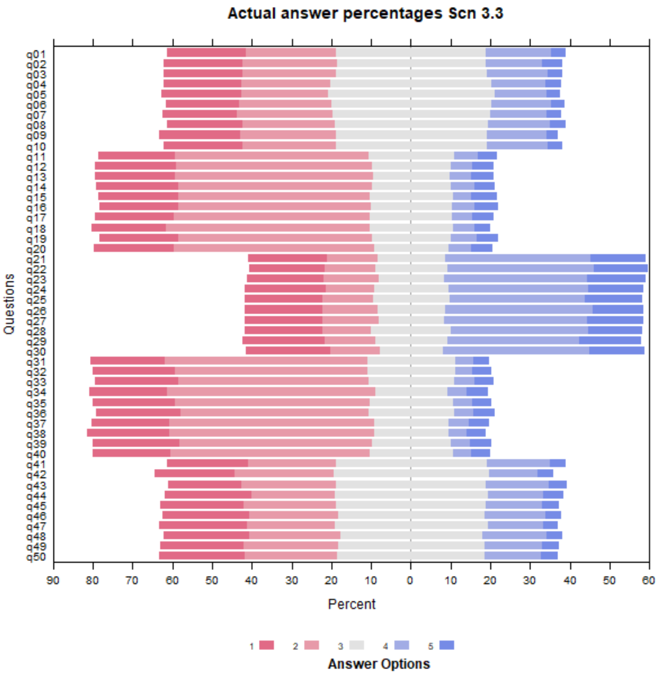


Figure 15: Scenario 3.3 with 5-point scale

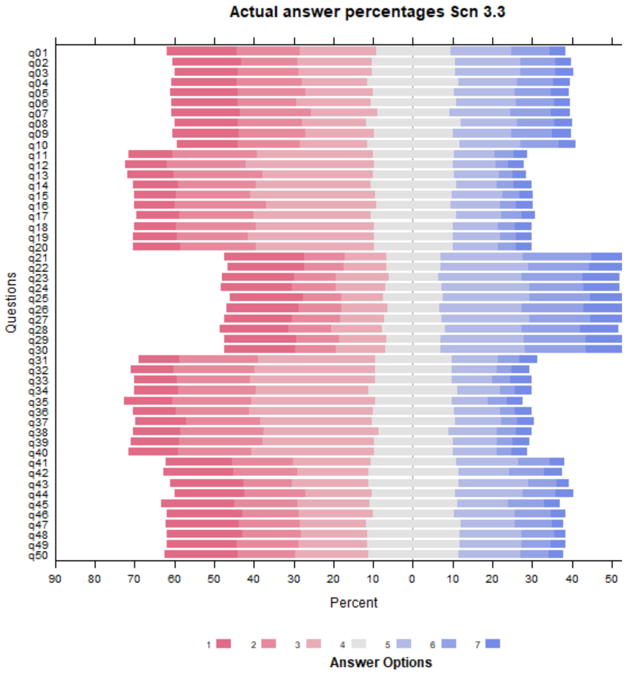


Figure 16: Scenario 3.3 with 7-point scale

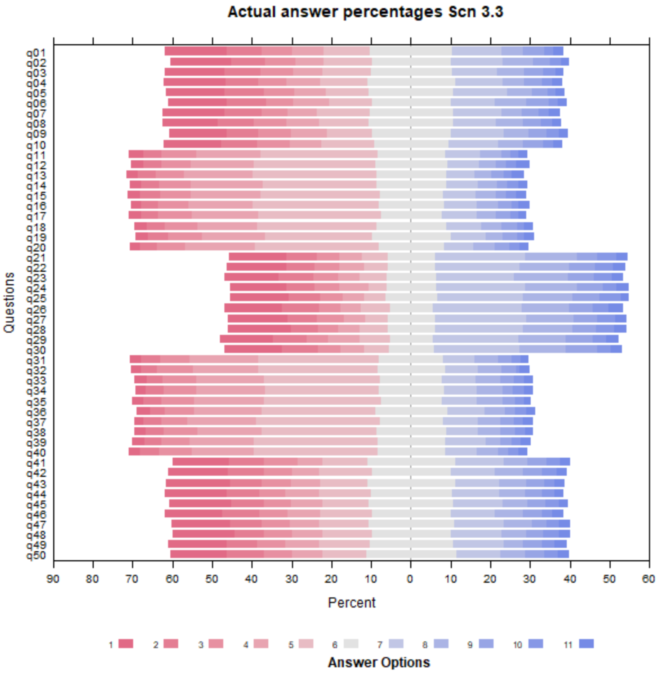


Figure 17: Scenario 3.3 with 11-point scale

The following figures show the scatterplots of the 5 invariant coordinates that are available in the context of average constructs, for the different scenarios and scatter pairs mentioned in the captions.

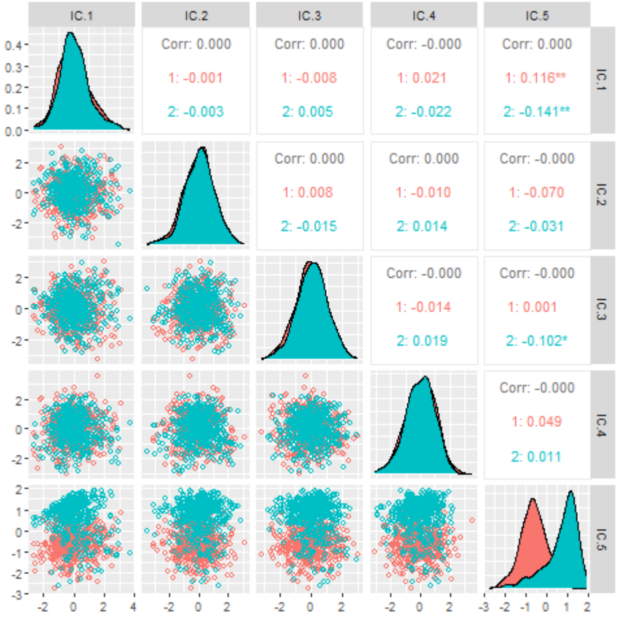


Figure 18: Scenario 2.2 with 11-point scale, $COV - COV4$

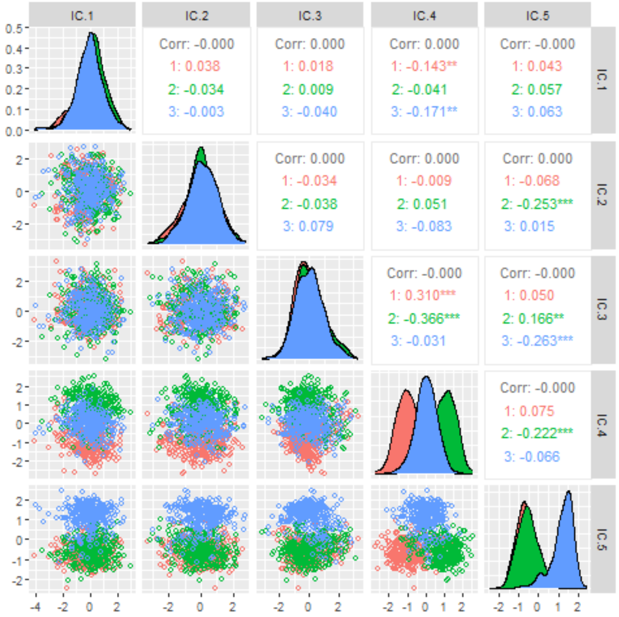


Figure 19: Scenario 2.3 with 11-point scale, $COV - COV4$

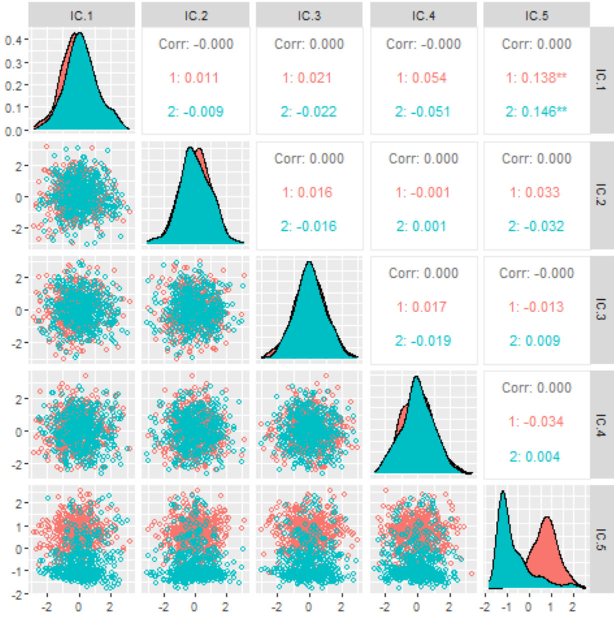


Figure 20: Scenario 3.2 with 11-point scale, $COV - COV4$

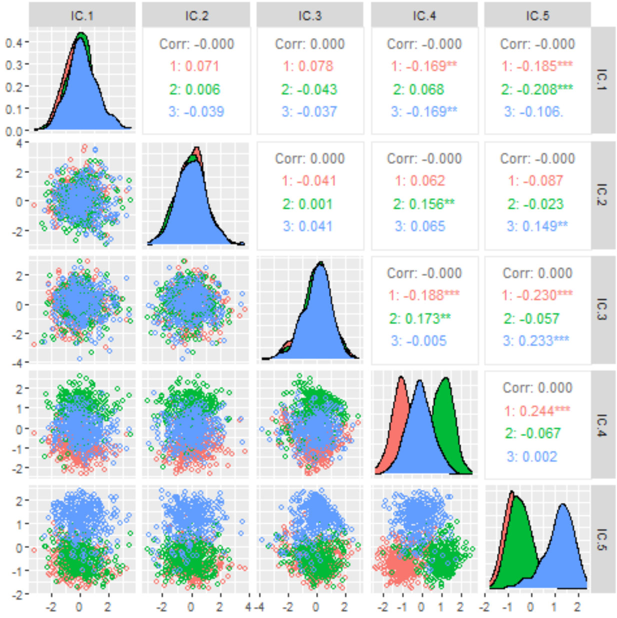


Figure 21: Scenario 3.3 with 11-point scale, $COV - COV4$

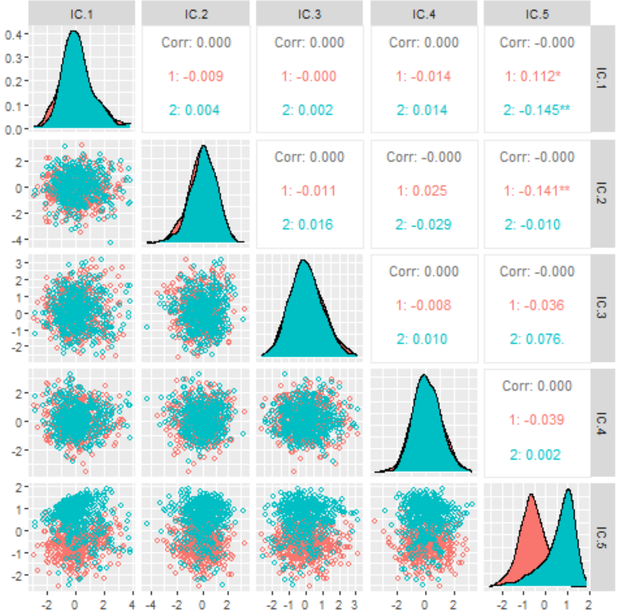


Figure 22: Scenario 2.2 with 11-point scale, $MCD - COV$

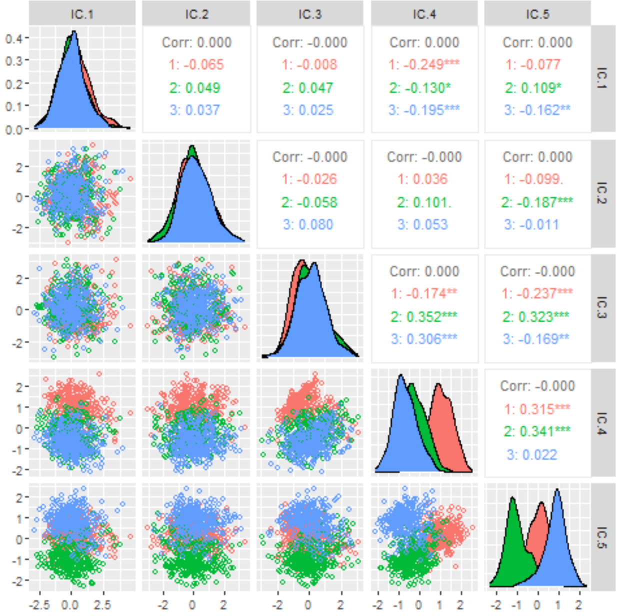


Figure 23: Scenario 2.3 with 11-point scale, $MCD - COV$

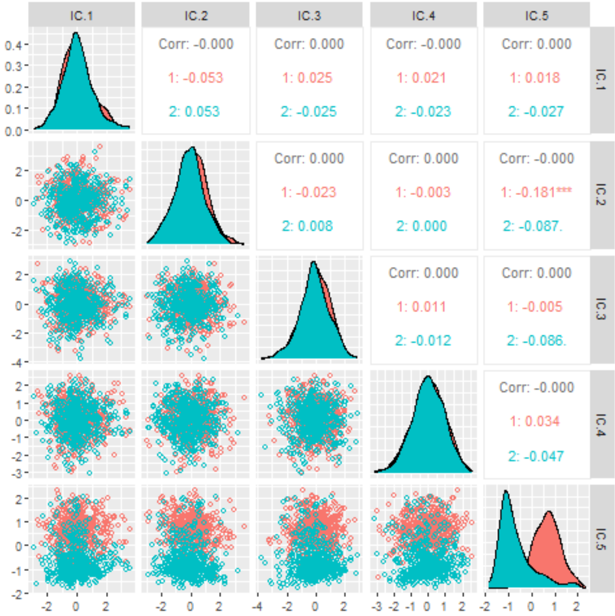


Figure 24: Scenario 3.2 with 11-point scale, $MCD - COV$

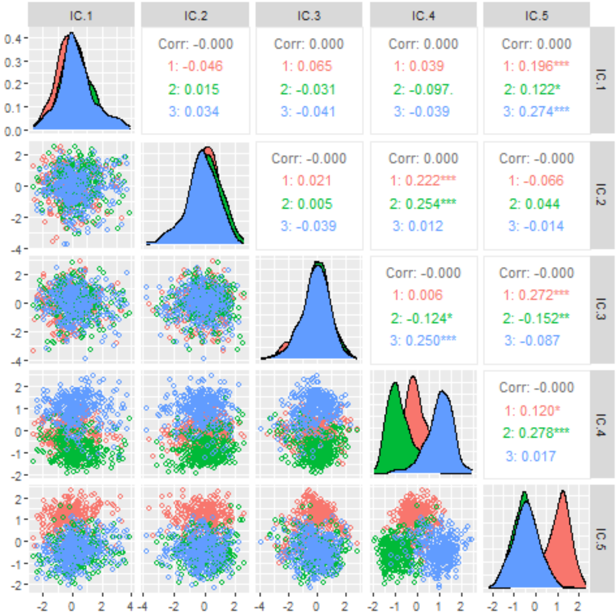


Figure 25: Scenario 3.3 with 11-point scale, $MCD - COV$

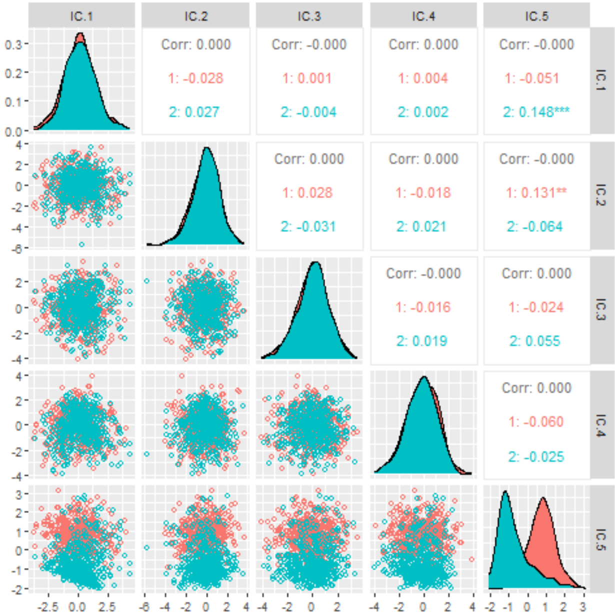


Figure 26: Scenario 2.2 with 11-point scale, $TM - COV$

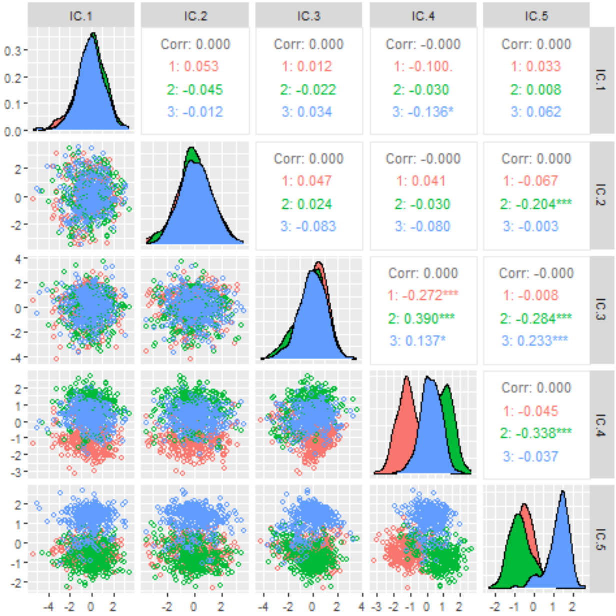


Figure 27: Scenario 2.3 with 11-point scale, $TM - COV$

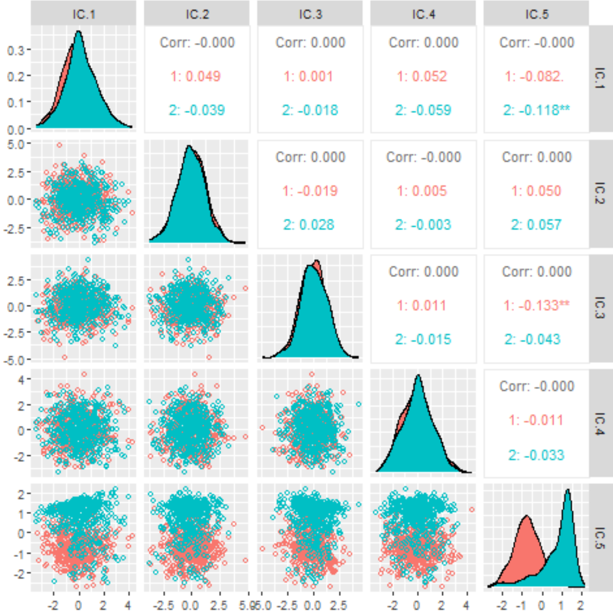


Figure 28: Scenario 3.2 with 11-point scale, $TM - COV$

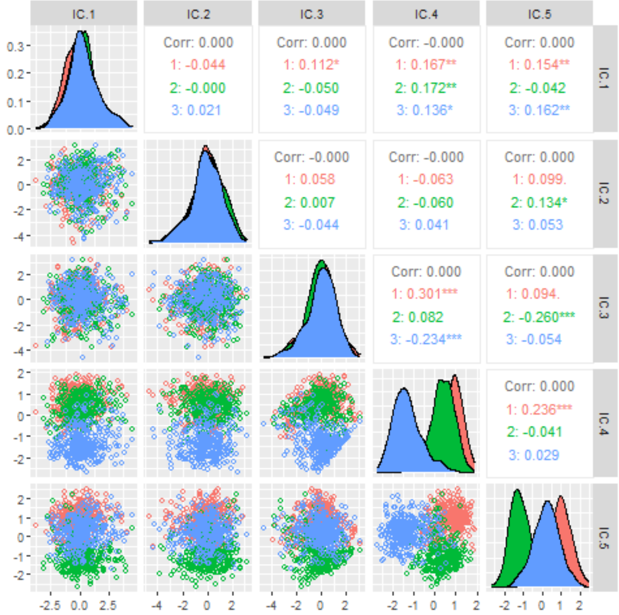


Figure 29: Scenario 3.3 with 11-point scale, $TM - COV$

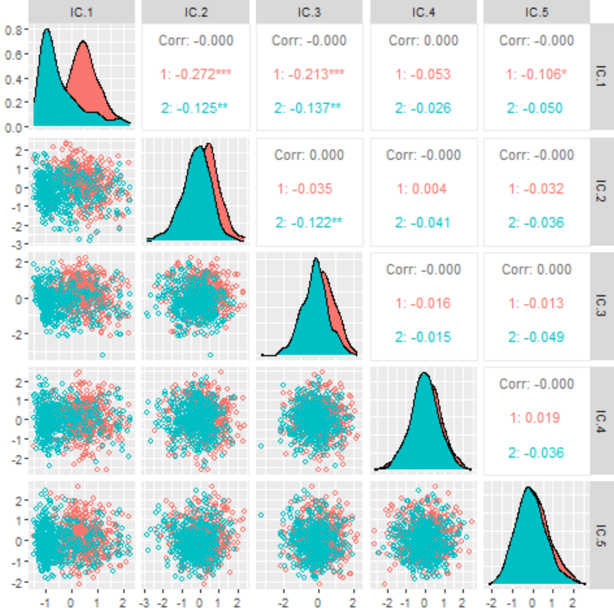


Figure 30: Scenario 2.2 with 11-point scale, $LCOV - COV$

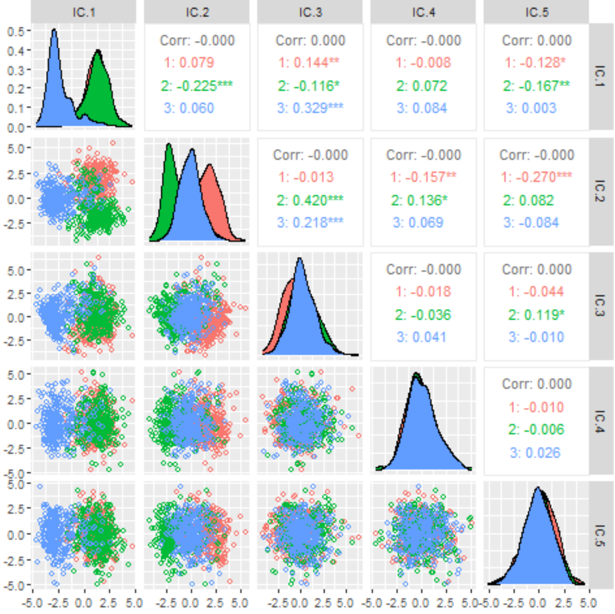


Figure 31: Scenario 2.3 with 11-point scale, $LCOV - COV$

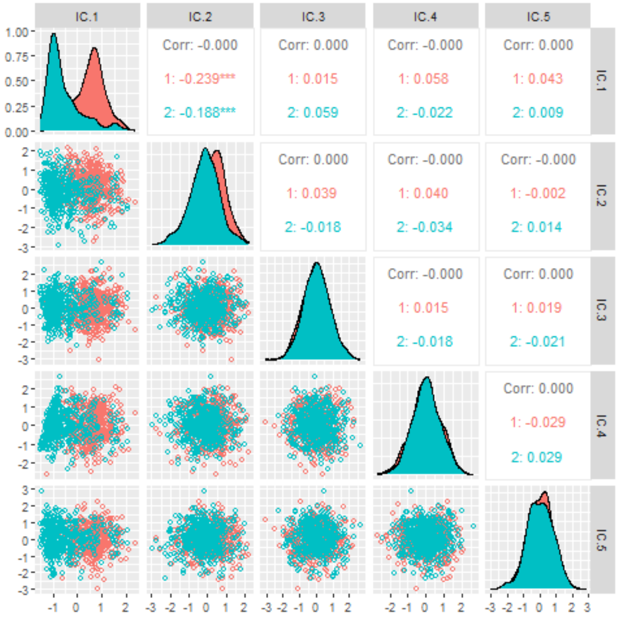


Figure 32: Scenario 3.2 with 11-point scale, $LCOV - COV$

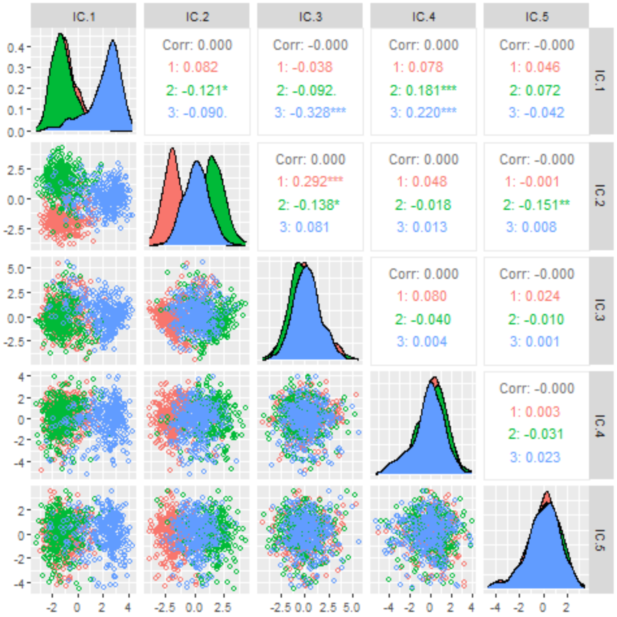


Figure 33: Scenario 3.3 with 11-point scale, $LCOV - COV$