Subspace Independent Component Analysis (SICA) A comparison of methods for cluster analysis in high dimensionality.

Rozema, Lude

Faculteit der Sociale Wetenschappen

Subspace Independent Component Analysis (SICA).

A comparison of methods for cluster analysis in high dimensionality.

Lude Jan Rozema

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: December 14th, 2021

Student number: X

Supervisor: Dr. Tom F. Wilderjans and Jeffrey Durieux, MSc.

Abstract

Clustering algorithms are important for data mining, and K-means is one of the most well-known clustering algorithms currently available. In cases in which data are high-dimensional, however, mere application of K-means to a data set may fail to uncover clusters due to presence of masking variables, the curse of dimensionality, and difficulties in interpretation of the obtained solution. A commonly used work-around is to apply dimension reduction to the data prior to performing cluster analysis, a practice called Tandem Analysis (TA). A vulnerability of TA is that the applied dimension reduction is not guaranteed to preserve cluster structure present in the original data, jeopardising the usefulness of subsequent cluster analysis. Multiple authors have provided algorithms that reduce dimensionality of a data set and perform cluster analysis on the reduced data, either in a sequential fashion or a simultaneous fashion, all aiming to find suitable low-dimensional representations of data while also keeping cluster structures intact. In this thesis, a novel approach to reducing dimensionality and performing cluster analysis on the low dimensional representation of the data - called SICA - is described and thoroughly tested in two systematically manipulated simulation studies and applied to three empirical data applications. Results show that SICA is a computationally efficient algorithm well able to extract components from the original data that preserve cluster structures, but that performance depends on characteristics of the data and the model of data generation. In addition, the correctness and validity of the clusterings obtained through SICA is high, although it does not always outperform currently available methods in this regard and is dependent on the same characteristics of the data and model generation as the other algorithms. Limitations and implications for future research are discussed.

Contents

1 Introduction

In various academic and applied fields, cluster analysis is a commonly used method for identifying and describing subgroups of subjects or individuals present in data. These subgroups, or clusters, are based on patterns of similarity in variable profiles and may shed light on important heterogeneity between subjects in the observed concept at hand. Traditional clustering algorithms such as K-means (MacQueen, 1967) use the full dimensionality of the data (i.e. all variables in the data) to find potential cluster structures in the data. Although the application of clustering algorithms to any full dataset is in line with the general purpose of cluster analysis, the efficacy of clustering algorithms may be compromised when the nature of the data is high-dimensional. This is due to several reasons.

First, in large datasets, some variables may not at all contribute to cluster structures or, worse, mask existing cluster structures present in the data (Milligan, 1980; Parsons, Haque, & Liu, 2004). When applying traditional clustering algorithms to such data, the presence of these noise or masking variables results in decreased performance (see, for example, the simulation conducted by Steinley & Brusco, 2008).

Second, as the number of dimensions increases for a given number of data points, the distance between each data point and its *nearest* neighbour increases, eventually approximating the distance between that same data point and its *most distant* neighbour. Since clustering algorithms typically apply some distance metric to quantify similarity between points, the effect of increasing dimensionality on mutual distance between data points in the full-dimensional space (the so called 'curse of dimensionality') is that distance metrics will not adequately quantify (dis)similarity between data points, jeopardising the accuracy of these algorithms (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999;

Steinbach, Ertöz, & Kumar, 2004).

Lastly, when cluster analysis has been performed, a researcher will typically try to interpret the obtained clusters substantively in the light of some research question. Although a two- or three-dimensional cluster structure is well-suited for interpretation, any representation of a cluster structure in a higher-dimensional space does not allow for an easy substantive interpretation. This latter drawback is not so much a statistical one as the former two, but is of large importance in the usefulness of cluster analysis nonetheless.

Several strategies have been proposed (and explored) to overcome these drawbacks. One consists of applying weights to the variables based on their importance in the underlying cluster structure, assigning higher weights to more important variables and lower ones to variables that contribute little to the cluster structure (see, for instance, Steinley & Brusco, 2008). The weights, in the example of Steinley and Brusco (2008) are based on the ratio of a variable's range to its variance and are thus defined a priori. A weight of 0 can be assigned to a variable if that variable does not contribute to the cluster structure at all, amounting to a type of variable selection. One typically wants to assign such a zero weight to noise and masking variables.

A second strategy consists of applying a clustering algorithm to a lower-dimensional representation of the original data that contains as much information as possible about the original data. This can be achieved by applying a data reduction technique to the data (e.g. Principal Component Analysis; PCA) and applying a clustering algorithm such as K-means to the first few (linear) projections of the original data provided in the *reduced* data. The procedure of applying PCA and consecutive K-means to the components is called *tandem analysis* (TA) and its intuitive appeal lies in the fact that PCA, in its first few principal components, explains as much variance as possible (formally by min-

imising the sum of squares between the original data and some superimposed linear projection making up the individual components), containing as much information as possible about the original data in a (much) lower dimensional projection.

Despite the fact that tandem analysis offers a clear strategy to overcome the drawbacks of cluster analysis in high dimensional settings, a well-known and warned against complication of this method is that the components extracted are not guaranteed to *preserve* any potentially present cluster structure in the original data (De Soete & Carroll, 1994; Arabie & Hubert, 1996). This is due to the fact that the individual components extracted by the PCA are extracted in such a way that the *variance* of each of the individual components is maximised, often resulting in more or less unimodal, Gaussian components. Components exhibiting cluster structure, in contrast, are expected to take on a more *bimodal or multimodal* shape (the clusters of (dis)similar observations making up the modes), but PCA is not guaranteed/able to extract components of such a shape.

Over time, multiple methods have been developed aimed at overcoming the difficulties in cluster analysis associated with high dimensionality beyond the suboptimal solution provided by tandem analysis. Notable examples are Reduced K-means (RKM; De Soete & Carroll, 1994), Factorial K-means (FKM; Vichi & Kiers, 2001), Subspace K-means (SKM; Timmerman, Ceulemans, De Roover, & Van Leeuwen, 2013) and Principal Cluster Axes Projection Pursuit (PCAPP; Steinley, Brusco, & Henson, 2012). These methods all apply component analysis or a similar projection method to reduce the dimensionality of the original data, in combination with a clustering algorithm. However, since all these methods (except PCAPP) capitalise on PCA as a means of dimension reduction, the aforementioned characteristics of their extracted components may well fail to return cluster structures embedded in subspaces of the original data. PCAPP,

instead, maximises an index known as the Clusterability Index (*CI*; Steinley et al., 2012) instead of variance, in order to obtain components of non-Gaussian shapes. As such, PCAPP explicitly searches for projections of the data with a high potential for clustering.

Another method that might be better equipped to return non-Gaussian structures embedded in a subspace of the original data that contain relevant clustering information is Independent Component Analysis (ICA; Comon, 1994; Hyvärinen & Kano, 2003), a method coming from the field of signal-processing. ICA is a multivariate latent variable model that searches for maximally *non-Gaussian* projections in multivariate data, which are mutually statistically independent. Applying ICA to a dataset containing a cluster structure in its subspace will yield projections that exhibit a non-Gaussian (e.g. bimodal or multimodal) structure, making it a suitable alternative to find the relevant clustering in the data.

ICA provides an attractive alternative to PCA in reducing dimensionality of data while also preserving potentially present cluster structures residing in the data subspace. However, ICA has the ambiguity that the order in which it estimates these components is free to vary without affecting the overall model outcome. In comparison, PCA estimates its components in the same order every time, as the components are ordered in terms of the amount of variance they explain in the data. ICA does not possess such a property and, as a consequence, application of ICA to a multivariate data set to find components exhibiting cluster structure may require visual inspection of each individual component or some other time-consuming endeavour. This is not a realistic strategy for large data sets, and so it is proposed to combine ICA with a metric assessing the multimodality of the returned projections called the Bimodality Coefficient (*BC*; Freeman & Dale, 2013) to return individual components exhibiting cluster

structures. This method is called Subspace-ICA (SICA; Durieux & Wilderjans, 2019) and will be central to this thesis.

Preliminary results show that SICA recovers cluster structures residing in a subspace of the full data well over multiple different conditions (e.g. number of clusters or amount of noise present in the data), returns components that exhibit more cluster structure than components extracted by traditionally applied methods and is computationally efficient (Durieux & Wilderjans, 2019). Furthermore, application of SICA to toy data reiterates these findings and demonstrates the performance of SICA in finding well-interpretable and well-separated clusters. Based on these findings, it can be hypothesised that SICA will outperform other (more traditional) methods in terms of finding relevant cluster structure in (a subspace of) high dimensional data. Therefore, the aim of this thesis is to subject SICA and related methods to a more stringent comparison by means of two simulation studies and an application to three realistic empirical data problems.

This thesis is structured as follows. In section 2, an elaborate description of all methods used in this thesis will be provided in terms of their inner workings and mathematical formalities. Section 3 contains the simulation studies and their results, and section 4 presents the application to two functional magnetic resonance imaging (fMRI) data sets and one data set containing measures of sensory processing sensitivity (SPS). Section 5 contains the discussion and limitations of this thesis, along with possible future directions for research.

2 Existing methods

In this section, an overview of currently existing methods that combine dimension reduction and clustering techniques in either a sequential or simultaneous fashion will be given.

2.1 *K*-means clustering

The K-means clustering algorithm (MacQueen, 1967; Forgey, 1965; Lloyd, 1982; Hartigan & Wong, 1979) is one of the most commonly used clustering algorithms currently available, and is central to all methods discussed in this chapter. K-means seeks to partition *N (i = 1, 2 ..., N)* objects into *K (k = 1, 2 ..., K)* homogeneous, hard clusters in which every subject belongs to one and only one cluster, such that $K \ll N$. It does so by minimising the following loss function:

$$Q(\mathbf{P}, \mathbf{C}) = ||\mathbf{X} - \mathbf{PC}||^2 \tag{1}$$

This amounts to minimising the within-cluster variance between the subjects of a cluster or the within-cluster distances between the subjects and their associated cluster centroid. In this equation, $\mathbf{X}$ denotes the $N \times P$ observed data matrix, $\mathbf{P}$ is an $N \times K$ binary indicator matrix in which, for each of 1 to N objects, the element i equals 1 if and only if that subject belongs to that cluster k. Lastly, the rows of the $K \times P$ centroid matrix $\mathbf{C}$ contain the centroid vectors for each of the K clusters. Estimation of the partitioning matrix $\mathbf{P}$ and the centroid matrix $\mathbf{C}$ can be done using an alternating least-squares type of algorithm which estimates $\mathbf{P}$ with respect to the current estimate of $\mathbf{C}$ after which it estimates $\mathbf{C}$ given the current $\mathbf{P}$. These steps are alternated repeatedly until convergence. The first step in this algorithm is using some initial estimate

of $\mathbf{P}$, which is obtained using random assignment of objects to the K clusters. Since these types of algorithms are sensitive to local minima, it is advised to use a multiple random start procedure in conjunction with K-means. Such a procedure runs the K-means algorithm several times, each time with a different (random) initialisation of $\mathbf{P}$. The solution with the lowest loss value encountered across all runs is selected as the final solution.

2.2 Tandem analysis

Tandem analysis, as described in the introduction, requires first reducing the dimensionality of an observed data set (using, for example, PCA) that preserves as much information as possible from the original data, after which some clustering algorithm is applied to the *reduced* data set. PCA projects an observed data set $\mathbf{X}$ to a space in which the individual components are orthogonal and explain as much variance as possible using the linear transformation $\mathbf{X}_r = \mathbf{X}\mathbf{E}_r$ where $\mathbf{E}_r$ is an orthonormal matrix (with r columns) containing the first r eigenvectors of the covariance matrix of $\mathbf{X}$ ordered by the strength of their accompanying eigenvalues as $\lambda_1 > \lambda_2 > ... > \lambda_r$ (Shlens, 2014). In practice, this linear transformation can be obtained by means of a singular value decomposition of $\mathbf{X}$ (for more information, see Shlens, 2014).

Although intuitively reasonable, there is no guarantee that the data reduction method applied to the full data preserves potential cluster structures embedded in it. This jeopardises the usefulness of tandem analysis in finding cluster structures in high-dimensional data.

2.3 Simultaneous dimension reduction and clustering

Unlike tandem analysis, in which dimension reduction and clustering take place in a sequential manner, multiple methods exist that combine dimension re-

duction and clustering in a *simultaneous* fashion. The gain of these methods over tandem analysis is that, by simultaneously reducing dimensionality and performing cluster analysis, potential cluster structures underlying the data are contained to a larger extent in the lower-dimensional representation, potentially improving performance of applying cluster analysis (De Soete & Carroll, 1994).

2.3.1 Reduced K-means

Reduced K-means (RKM; De Soete & Carroll, 1994) aims to perform dimension reduction on the full data matrix $\mathbf{X}$ while simultaneously clustering objects into a predefined number of hard clusters. The formal loss function it aims to minimise is much the same as Equation 1 but with the constraint that the rank of the centroid matrix $\mathbf{C}$ is lower than that of the full data matrix $\mathbf{X}$ (De Soete & Carroll, 1994). In matrix notation (based on Timmerman, Ceulemans, Kiers, & Vichi, 2010; adapted to match current notation):

$$Q(\mathbf{P}, \mathbf{C}, \mathbf{A}) = ||\mathbf{X} - \mathbf{PCA}||^2 \tag{2}$$

Note RKM achieves this lower-rank representation using an R-truncated singular value decomposition of the data matrix $\mathbf{X}$ in which only the first R singular values and left and right singular vectors are used (De Soete & Carroll, 1994). The value of the number R is defined by the user beforehand. All matrices in Equation 2 are the same as in Equation 1, except for the $(M \times R)$ orthonormal loading matrix $\mathbf{A}$.

De Soete and Carroll (1994) provide an alternating least-squares algorithm that minizes Equation 2 alternating between minimising $\mathbf{P}$ given the current estimate of $\mathbf{C}$ and minimising $\mathbf{C}$ given the updated estimate of $\mathbf{P}$ until con-

vergence is reached (De Soete & Carroll, 1994). This results in a clustering in which the centroid vectors that make up the rows of $\mathbf{C}$ are restricted to lie in an R-dimensional subspace of the full data matrix $\mathbf{X}$ (De Soete & Carroll, 1994). As in K-means, a multiple random start procedure is adopted to prevent the algorithm from converging into a single local minimum.

2.3.2 Factorial K-means

Vichi and Kiers (2001) pointed out that the loss-function that is minimised in RKM increases considerably when the observed data in $\mathbf{X}$ contain large directions of variance *orthogonal* to the subspace in which the cluster centroids in $\mathbf{C}$ reside. Factorial K-means (FKM; Vichi & Kiers, 2001), instead, minimises a different loss-function in which the sum of squares are calculated for each cluster using the cluster centroids in reduced space *and* orthogonal projections of the subject data points in reduced space. This is contrary to reduced K-means, in which the sum of squares are calculated between the cluster centroids in reduced space and the subject data points in the *full* data space (De Soete & Carroll, 1994). This reduces the magnitude of the loss-function and is more in line with the general thought of clustering in reduced dimensionality; that cluster structures and the subjects that shape them lie in a subspace of the original data. The formal model underlying FKM is given by:

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

in which $\mathbf{E}$ is a $(N \times K)$ Gaussian error term and $\mathbf{A}$ $(K \times M)$ an orthonormal matrix defining the subspace of $\mathbf{X}$ (Vichi & Kiers, 2001). $\mathbf{P}$ and $\mathbf{C}$, in turn, are equivalent to their RKM-counterparts. The formal loss-function minimised to estimate the FKM parameters is:

$$Q(\mathbf{P}, \mathbf{C}, \mathbf{A}) = ||\mathbf{XAA}' - \mathbf{PCA}'||^2 = ||\mathbf{XA} - \mathbf{PC}||^2 \tag{4}$$

Since the optimal cluster centroid matrix can be written as $\mathbf{C} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{XA}$, the loss-function can be further simplified to:

$$Q(\mathbf{P}, \mathbf{A}) = ||\mathbf{XA} - \mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{XA}||^2 \tag{5}$$

which is then estimated using a multiple random-start alternating least-squares algorithm in which $\mathbf{P}$ and $\mathbf{A}$ are updated alternatingly (for more information, see Vichi & Kiers, 2001 and Timmerman et al., 2010).

2.3.3 Subspace *K*-means

The last method that capitalises on PCA in simultaneously reducing dimensionality and performing cluster analysis is Subspace K-means (SKM; Timmerman et al., 2013). Like RKM and FKM, SKM models cluster centroids to reside in a subspace of the original data and performs K-means clustering on this reduced space. In addition, however, SKM allows to formally model the *residuals* of these clusters in a subspace of the original data as well, using so called within-cluster components. These components allow to describe the differences between subjects in the same cluster and, as a result, explain more variance than RKM or FKM. Formally, the SKM model is defined as

$$\mathbf{x}_i = \sum_{k=1}^{K} p_{ik}({\mathbf{f}_b}^k {\mathbf{A}_b}' + {\mathbf{f}_\mathbf{w}}_i^k + {\mathbf{A}_\mathbf{w}}^{k\prime}) + \mathbf{e}_i^k \tag{6}$$

Here, $\mathbf{x}_i$ denotes (mean centered) row i from $\mathbf{X}$, p_{ik} a binary scalar specifying to which cluster k subject i belongs collected in the *(N × K)* binary partitioning matrix $\mathbf{P}$, ${\mathbf{f}_b}^k$ a $1 \times Q_b$ vector containing *between*-component scores for cluster k ($k = 1, 2 ..., K$), ${\mathbf{A}_b}'$ a ($P \times Q_b$) matrix containing the between-cluster loadings in which P is the number of observed variables in the full data matrix $\mathbf{X}$ and Q_b is the number of between-cluster components. $\mathbf{f}_{\mathbf{w}i}^k$ is the ($1 \times Q_w$) within-component vector for individual i in cluster k collected in the (N × Q_w) within-component matrix ${\mathbf{F}_\mathbf{w}}^k$ for cluster k, ${\mathbf{A}_\mathbf{w}}^{k\prime}$ the ($P \times Q_w$) *within*-component loading matrix for cluster k (Q_w being the number of within-cluster components, kept the same for all clusters). Lastly, the ($1 \times P$) vector $\mathbf{e}_i^k$ contains the residuals for subject i in cluster k.

For any chosen number of clusters K, between-cluster components Q_b and within-cluster components Q_w, SKM optimises the following loss-function:

$$Q(\mathbf{P}, \mathbf{F_b}, \mathbf{A_b}, {\mathbf{F}_\mathbf{w}}^k, {\mathbf{A}_\mathbf{w}}^k) = \sum_{i=1}^{I} \left\| \mathbf{x}_i - \sum_{k=1}^{K} p_{ik}({\mathbf{f_b}}^k {\mathbf{A_b}}' + \mathbf{f}_{\mathbf{w}i}^k {\mathbf{A}_\mathbf{w}}^{k\prime}) \right\|^2 \quad (7)$$

Where the individual matrices and vectors are those as described in Equation 6. This loss-function is then optimised using an alternating least-squares algorithm with multiple-start procedure to prevent the algorithm from converging into a local minimum (Timmerman et al., 2013).

2.4 Principal Cluster Axes Projection Pursuit (PCAPP)

The methods described so far all rely to some extent on PCA (either in simultaneous or subsequent fashion) in combination with K-means clustering to find cluster structures embedded in large multivariate data. A pitfall of PCA-based

methods is that it extracts components by maximising variance for each of the components. As a consequence, the structure of the first few extracted components is often Gaussian. Gaussian components generally contain less clustering-potential than bimodal or multimodal components do (Friedman & Tukey, 1974; Hubert & Arabie, 1985; Steinley et al., 2012). Therefore, when aiming to find individual components that preserve and contain cluster structures in data, it is useful to move away from PCA as a means of dimension reduction.

A method that does *not* apply PCA as a means of data reduction is Principal Cluster Axes Projection Pursuit (PCAPP; Steinley et al., 2012). PCAPP is derived from a technique called Projection Pursuit (PP; Friedman & Tukey, 1974) which aims at finding individual projections of data - given by an optimal vector $\mathbf{w}_i$ - that maximises some user-defined index I of the projected data. This index can be chosen to suit any data analysis problem and is, therefore, not a single given metric. For any selected univariate index, however, it holds that:

$$\mathbf{z} = I(\mathbf{X}\mathbf{w}) \tag{8}$$

In which $\mathbf{z}$ is the individual projection obtained from the full data based on the index I and the optimal vector $\mathbf{w}$. Note that this index can also be defined such that all $\mathbf{w}_i$ in $\mathbf{W}$ are mutually orthogonal and explain as much variance as possible, in which case PP is equivalent to PCA. Formally, this would require maximizing $\mathbf{z} = var(\mathbf{X}\mathbf{w})$, subject to the constraint that all components are orthogonal.

The flexibility of being able to choose any index raises the question which index should be chosen to obtain cluster structures embedded in high dimensions. Steinley et al. (2012), along with other authors (Friedman & Tukey, 1974), ar-

gue that, in order to find cluster structures embedded in high dimensions, the aim should be to find projections that deviate from Gaussianity since these give more insight in potential cluster structures present in the data than Gaussian projections do (Steinley et al., 2012). In order to do so, PCAPP applies a measure called the clusterability index (CI; Steinley et al., 2012) which, for any projection, is given by:

$$CI = \frac{12 \times \sigma^2(x)}{(r(x))^2} \tag{9}$$

Here, $\sigma^2(x)$ indicates the variance and $r(x)$ the range of some variable or projection x. The CI has the attractive property of being both affine invariant (i.e., its projections are invariant to linear transformations) and robust to outliers (Steinley et al., 2012). PCAPP applies CI by estimating unit vectors $\mathbf{w}_i$ in a deflating fashion (i.e., each time removing the projection from the data and estimating the next projection on the deflated data) in such a way that the CI for each of the components is maximised. Once this has been achieved, PCAPP ensures all unit vectors $\mathbf{w}_i$ in $\mathbf{W}$ to be mutually orthogonal using Gram-Schmidt orthogonalisation (for more information, see Steinley et al., 2012).

2.5 Subspace Independent component analysis

2.5.1 Independent Component Analysis

ICA is a method from the field of blind source separation and constitutes a latent variable model aimed at finding interesting projections of multivariate data, much like PCAPP. Specifically, it finds a linear transformation of the data matrix $\mathbf{X}$ such that the extracted components are non-Gaussian and statistically independent (Comon, 1994; Hyvärinen & Kano, 2003). Application of PCA on Gaussian variables guarantees the extracted components to be uncorrelated. For

non-Gaussian variables, however, PCA will not suffice in extracting independent components (as independence is a more strict property than uncorrelatedness), and the additional constraint of statistical independence is required to guarantee this property of the components.

ICA is a specific case of the more general framework of Invariant Coordinate Selection (ICS; Tyler, Critchley, Dümbgen, & Oja, 2009). ICS can be used to find cluster directions in multivariate data and requires the definition of two scatter-matrices that describe the dispersion in the data matrix $\mathbf{X}$, which are then compared in relation to their eigenvalues and eigenvectors. More specifically, ICA is a case of ICS in which *(1)* all components in $\mathbf{S}$ are mutually independent, *(2)* at most one of the components in $\mathbf{S}$ is *not* symmetrically distributed around 0 and *(3)* all eigenvalues of $\mathbf{X}$ are distinct, meaning non-identical.

ICA assumes that N (x_1, x_2, ..., x_n) observed random variables are, in fact, linear transformations of m latent variables (s_1, s_2, ..., s_m), called sources. In matrix notation:

$$\mathbf{X} = \mathbf{SA} \tag{10}$$

Here, $\mathbf{X}$ is the $(N \times P)$ data matrix, $\mathbf{S}$ $(N \times P)$ the matrix containing the independent components in its columns (also called the source matrix), and $\mathbf{A}$ the mixing matrix containing the parameters that transform $\mathbf{S}$ to the observed data $\mathbf{X}$. $\mathbf{A}$ is often assumed to be a square matrix, meaning that the amount of latent sources M is equal to the number of observed variables N (its dimensions, then, would be $P \times P$). This, however, is merely an assumption, and does not have to be the case; the number of latent sources can differ from the number of observed variables.

In order to estimate the independent components in $\mathbf{S}$, the inverse of $\mathbf{A}$ (i.e. $\mathbf{A}^{-1} = \mathbf{W}$) is estimated to invert the linear system that yields $\mathbf{X}$. Since both $\mathbf{A}$ and $\mathbf{S}$ are unknown, however, estimating $\mathbf{W}$ is done based on the non-Gaussian properties of the components in $\mathbf{S}$. FastICA is a computationally efficient implementation of ICA that estimates $\mathbf{W}$ by maximising the negative entropy (*negentropy*) of the individual components in $\mathbf{S}$ (Hyvarinen, 1999). Negentropy is a metric quantifying the extent to which a given variable deviates from a Gaussian distribution, having a value of zero for Gaussian distributed variables and increasing for components that deviate from a Gaussian variable. Maximising the negentropy of the components in $\mathbf{S}$, therefore, allows to estimate components that are statistically independent and non-Gaussian. Formally, denoting J an index for non-Gaussianity (e.g. negentropy), all $\mathbf{w}_i$ in $\mathbf{W}$ are selected by maximising the index of non-Gaussianity of the projected data as:

$$\hat{\mathbf{S}}_i = J(\mathbf{X}\mathbf{w}_i) \tag{11}$$

A preprocessing step applied in FastICA to aid in extracting these components is that data are *whitened* before ICA is applied. Whitening the data first removes all linear dependencies among the data by means of a PCA, after which it normalises the variance to a value of 1. After whitening the data, the optimal loading matrix $\mathbf{A}$ is known to be a rotation matrix, which is obtained such that all latent sources in $\mathbf{S}$ contain maximal negentropy. This reduces the number of parameters that needs to be estimated and increases the performance of the algorithm (Hyvarinen, Karhunen, & Oja, 2001).

Like PCA, ICA can be formulated as a specific case of PP in which the maximised index is the negentropy. This relation only holds, however, when

the assumptions of non-Gaussianity and statistical independence of the ICA model hold. PP does not require these assumptions and so, if the assumptions do not hold, application of ICA results in PP-directions of the data (Durieux & Wilderjans, 2019). The assumptions underlying ICA also dictate that, in the case in which there is more than one component in $\mathbf{S}$ that is Gaussian as opposed to non-Gaussian, there is indeterminacy in $\mathbf{A}$ (Tyler et al., 2009), and $\mathbf{S}$ can be only defined up to permutation and scaling (Tyler et al., 2009). A final ambiguity in the ICA model is that its components can change sign due to a change in orientation of the independent components in regards to the origin of the data space caused by a scaling of -1 (Shlens, 2014).

2.5.2 Previous applications of ICA in cluster analysis

The general ICA model has been applied for clustering purposes before and this paragraph will briefly describe these applications. One application of ICA in cluster-analysis seeks to find a single most non-Gaussian projection of the data by minimising the kurtosis of the data and visually inspecting the extracted latent sources for clusters (Bugrien & Kent, 2009). This method utilises the property of the ICA model that it extracts non-Gaussian components that may show interesting cluster structures. In the field of genomics, ICA has been applied as well. ICA has been used to extract latent features underlying RNA-sequence data and has been tested in combination with multiple different clustering procedures, including K-means (Feng et al., 2020). ICA has also been applied in RNA-sequencing data in which data were longitudinal (Nascimento et al., 2017). In the latter study, ICA was used to obtain individual, independent variables from the dependent temporal data and the extracted latent variables were subjected to K-means and hierarchical clustering analysis. These applications demonstrate that ICA can be used as a method of achieving reduced dimensionality as well as extracting statistically independent components from the

data. Lastly, ICA has been applied to data sets normally used in deep-learning based clustering methods to compare performance on such data. The appeal of ICA in this setting is that it is less complex than deep-learning based clustering methods but may provide comparable performance to these more complex methods while aiding in interpretation. Furthermore, ICA is less computationally taxing due to the absence of hyperparameters that require the tuning for deep-learning based methods to show good performance. Gultepe and Makrehchi (2018) showed that this is, indeed, the case as applying ICA in combination with spectral graph clustering performed about as well as much more complex methods based on deep-learning autoencoders.

2.5.3 Subspace ICA

As mentioned, the estimated components in $\mathbf{S}$ can be estimated up to a permutation meaning that they will, generally, not be returned in the same order (i.e., the order of the components is random and may differ when re-analysing the data with ICA). This becomes clear once the permutation matrix $\mathbf{O}$ *(*$P \times P$*)* and its inverse is added to the ICA model:

$$\mathbf{X} = \mathbf{SO}^{-1}\mathbf{OA} \tag{12}$$

This does not change the observed data in $\mathbf{X}$, and $\mathbf{SO}^{-1}$ now contains the latent sources ordered by $\mathbf{O}$, and $\mathbf{OA}$ the associated mixing matrix of this representation, with the mixing vectors ordered in the same way as the sources in $\mathbf{SO}^{-1}$. In order to become a unique ordering of components and retrieve components that show cluster structure, Subspace-ICA (SICA, Durieux & Wilderjans, 2019) utilises the Bimodality Criterion (*BC*, Freeman & Dale, 2013; SAS, 2004) to quantify the extent to which the components extracted by ICA demonstrate

cluster structures. BC, here, is based on the skewness and kurtosis of a variable and is given by:

$$BC = \frac{x_3^2 + 1}{x_4 + \frac{3(n-1)^2}{(n-2)(n-3)}} \tag{13}$$

Here, x_3 and x_4 indicate the skewness and kurtosis of some variable x, and n denotes the number of observations. BC-values vary from 0 to 1 and values larger than .555 are indicative of multimodality (SAS, 2004). SICA orders the obtained independent components by the magnitude of its univariate BC values and performs clustering on a predefined number of components herewith selecting the components with the largest BC values (irrespective their BC value, i.e. above .555 or not). Preliminary results suggest that SICA, by combining fastICA and BC, returns cluster structures well (Durieux & Wilderjans, 2019). Subsequent application of K-means to the extracted components can disclose these clusters and allow further interpretation. In the following simulations, the performance of SICA will be compared to the performance of the related methods described earlier in this section.

3 Simulation

In this section, SICA will be compared to the previously described other methods in terms of clustering performance by generating artificial and systematically manipulated data and applying these methods to these data. The generation of these data is conducted in two different ways, both of which will be described in a detailed fashion in their respective sections.

3.1 Design and evaluation

For both simulations, the design of Steinley et al. (2012) will be replicated, meaning the following factors will be manipulated:

- The number of clusters present in the data at three levels: 4, 6 and 8.
- The number of variables defining the subspace in which the clusters reside at four levels: 4, 6, 8 and 12.
- The number of masking variables obscuring the cluster structures in the data at four levels: 4, 6, 8 and 12.
- The way in which the subjects were divided over the clusters at three levels: equal, 10% and 60%. In the first condition (equal), subjects are distributed equally over the K clusters. In the second condition (10%), there is one small cluster containing 10% of subjects whereas the remaining subjects are distributed equally over the other K-1 clusters and in the third condition (60%) there is one large cluster containing 40 percent of subjects whereas the remaining subjects are distributed equally over the other K-1 clusters (Steinley et al., 2012).

For all analyses, the number of within-cluster components Q_w for SKM was fixed at 2 and the number of subjects divided over the clusters N was fixed

at 200 (a common number of subjects in cluster analysis; Steinley, 2003). The number of variables ranged from 8 (4 components + 4 masking variables) to 24 (12 components + 12 masking variables). For each level of this factorial design, ten replications were generated, totalling 3 (number of clusters) $\times$ 4 (number of components making up the subspace) $\times$ 4 (number of noise variables) $\times$ 3 (division of subjects over clusters) $\times$ 10 (replications) = 1440 data sets. Each simulated dataset was subjected to the six methods mentioned in Section 2 (TA, RKM, FKM, SKM, PCAPP and SICA), using the true number of clusters and components used to generate the data.

Evaluation of clustering performance and extraction of cluster structures in the data will, for both simulations, be conducted using the following metrics. First, clustering *accuracy* (i.e., the extent to which the obtained clustering matches the actual clustering) will be described using the Adjusted Rand-Index (ARI, Hubert & Arabie, 1985). The ARI compares the accuracy of the obtained clustering to the true clustering underlying the data and, in doing so, accounts for chance-level agreement between the two. It has a maximum of 1 (Yeung & Ruzzo, 2001), where 1 indicates perfect cluster recovery, 0 indicates recovery at chance level and negative values indicate erroneous cluster recovery. Second, the extent to which the obtained clusters demonstrate dense, compact clusters is evaluated using the Silhouette Index (SI; Rousseeuw, 1987). The SI compares within-cluster variance to the distance between the different clusters (i.e., between-cluster variance) and is bounded between -1 (indicating spread out, incoherent clusters) and 1 (indicating very dense, compact clusters). Milligan and Cooper (1985) and Arbelaitz, Gurrutxaga, Muguerza, Pérez, and Perona (2013) found that the SI describes cluster-properties well.

The extent to which each of the methods returns non-Gaussian subspaces will, for both simulations, be evaluated by applying *BC* (see Section 2.5.3;

Freeman & Dale, 2013) and the Dip-test (Hartigan & Hartigan, 1985) to each of the individual components/features extracted by the methods. The Dip test allows to statistically test whether a distribution is unimodal or multimodal (Hartigan & Hartigan, 1985). Components exhibiting multimodality with statistical significance lower than $\alpha = 0.05$ will be considered multimodal. In order to adequately discuss and compare the extent to which each of the 1440 data sets contained non-Gaussian components, the median BC-value and Dip-test p-value of the extracted components will be reported. Note that the median was reported because this metric is less sensitive to the skewed distribution of the BC-value and Dip-test p-values than the mean would be.

Lastly, computation time of each of the algorithms will be measured and evaluated to provide a metric of computational efficiency.

Software implementation for the described algorithms is the same for both simulation studies. For RKM and FKM, the 'clustrd' package version 1.3.7-2 (D'Enza, Markos, van de Velden, & Markos, 2016) was used whereas for KM and PCA, built-in functions were used in R-software. SKM and PCAPP code were custom made and is added in Appendix A.

It is expected that, on average, SICA is better than the other methods at identifying multimodal components, but that this deteriorates as the number of noise variables and the number of clusters present in the data increases (Durieux & Wilderjans, 2019; Steinley et al., 2012). In terms of goodness of recovery, it is expected that SICA is, on average, well able to recover cluster structure embedded in the data but that this is attenuated as the number of noise variables increases (Durieux & Wilderjans, 2019) and that this is also the fact as the number of clusters present in the data increases (Steinley et al., 2012). This will be tested by conducting a mixed analysis of variance (ANOVA) in which the methods make up the within-factor and the manipulated factors make up

the between-factors. These analyses were carried out using the *ez* *R*-software package (Lawrence, 2016).

3.2 Simulation 1

3.2.1 Data generation procedure

In this simulation, ICA and the related methods described so far will be subjected to the factorial design conducted by Steinley (2012), which is based on the procedure outlined by Milligan (1985) and implemented in an *R*-software package called *clusterGeneration* by Qiu and Joe (2006a) and Qiu, Joe, and Qiu (2006). The procedure in Milligan (1985) ensures that each of the generated clusters is separated on some subspace in the data, making for non-overlapping clusters. In addition, *clusterGeneration* allows to specify the extent to which the generated clusters are separated (Qiu & Joe, 2006a; Qiu et al., 2006; Qiu & Joe, 2006b) and allows to generate a full factorial design with multiple replications of data sets such as the design specified in this thesis. The index of separation was, for all designs, fixed at 0.3. Figure 1 illustrates the influence of this index, and shows an example of four clusters rendered on two dimensions with a separation index of 0.3.

3.2.2 Results simulation 1

Goodness of recovery Table 1 shows the average ARI values overall and per method and factor level, alongside the accompanying standard deviations. As can be seen from the bottom row, RKM yields the highest ARI values for all factor levels. For all factor levels, SICA clearly performs substantially worse than RKM, is in the same range as TA and SKM and outperforms FKM and PCAPP. ANOVA on the ARI-values was conducted to find main effects and interaction effects of the manipulated factor levels on goodness of cluster recov-

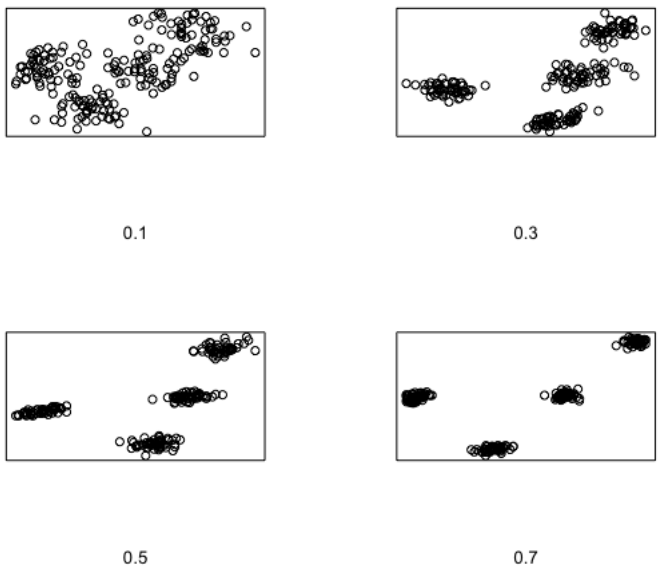

Figure 1. Illustration of the separation value for four clusters and two dimensions for separation values of 0.1, 0.3, 0.5 and 0.7. For this simulation, this index was fixed at 0.3, indicated by the upper right plot.

ery. Only effects that are significant beyond the standard $\alpha = .05$ level and have a generalised eta-squared measure (η^2_G) larger than 0.15 (Bakeman, 2005) will be reported. Multiple main- and interaction effects were found, of which the main effects will be discussed first. No higher-order interactions were deemed sufficiently interesting (based on η^2_G higher than 0.30) to discuss.

As expected based on the results in Table 1, there were significant differences between methods in terms of their ARI-values ($\eta^2_G = 0.56$; $F(5, 600) = 292.14$, $p < .001$). RKM shows the best overall average recovery as the average ARI value over all 1440 data sets is 0.86. In addition, SKM shows good overall performance with an overall ARI value of 0.71 and SICA follows with an overall average ARI value of 0.67.

Table 1

Average ARI values and standard deviations, overall and per method and factor level for the first simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.77	0.22	0.76	0.24	0.96	0.10	0.40	0.31	0.80	0.12	0.50	0.24	0.70	0.29
	6	0.62	0.22	0.62	0.30	0.86	0.24	0.69	0.28	0.71	0.24	0.46	0.23	0.66	0.28
	8	0.61	0.24	0.49	0.29	0.76	0.29	0.67	0.28	0.62	0.28	0.42	0.23	0.59	0.29
# Components	4	0.57	0.26	0.30	0.18	0.62	0.32	0.47	0.27	0.49	0.27	0.30	0.20	0.46	0.28
	6	0.68	0.24	0.53	0.25	0.89	0.19	0.62	0.32	0.74	0.19	0.41	0.21	0.64	0.28
	8	0.72	0.22	0.76	0.20	0.97	0.07	0.66	0.35	0.81	0.14	0.51	0.21	0.74	0.26
	12	0.72	0.21	0.89	0.13	0.96	0.08	0.59	0.30	0.79	0.14	0.63	0.20	0.76	0.23
# Noise variables	4	0.82	0.18	0.75	0.28	0.91	0.17	0.75	0.24	0.74	0.20	0.64	0.20	0.77	0.23
	6	0.74	0.20	0.66	0.30	0.88	0.22	0.68	0.28	0.72	0.22	0.51	0.21	0.70	0.27
	8	0.66	0.21	0.59	0.29	0.86	0.24	0.57	0.31	0.70	0.23	0.42	0.21	0.63	0.28
	12	0.46	0.20	0.48	0.26	0.79	0.29	0.35	0.28	0.67	0.26	0.28	0.17	0.51	0.30
Division	Equal	0.74	0.24	0.64	0.30	0.89	0.22	0.65	0.32	0.79	0.22	0.49	0.24	0.70	0.28
	Large	0.58	0.23	0.60	0.30	0.82	0.26	0.47	0.27	0.59	0.20	0.43	0.23	0.59	0.28
	Small	0.69	0.23	0.62	0.29	0.87	0.23	0.64	0.33	0.74	0.21	0.46	0.24	0.67	0.29
Overall		0.67	0.24	0.62	0.30	0.86	0.24	0.59	0.32	0.71	0.23	0.46	0.24	0.65	0.29

Note: Highest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

Regarding the manipulated factors, all factors except the number of clusters present in the data demonstrated to be of substantial main influence. Cluster recovery increases when the number of components increases ($\eta^2_G = 0.48$, $F(1, 120) = 230.56$, $p < .001$). As can be seen from Table 1 in conjunction with the upper right plot in Figure 2, all methods showed increased goodness of cluster recovery when the number of components increases. The nature of this increase, however, differed per method as there was a significant and substantial interaction between method and the number of components ($\eta^2_G = 0.23$; $F(5, 600 = 67.99$, $p < .001$). Specifically, the upper right plot in Figure 2 shows that this increase is linear for PCAPP and TA, whereas for the other methods, this trend decreases in steepness as the number of components increases. Durieux and Wilderjans (2019) and Steinley et al. (2012), too, found that goodness of recovery *increased* when the number of components making up the cluster space increased, and the current results reiterate such findings.

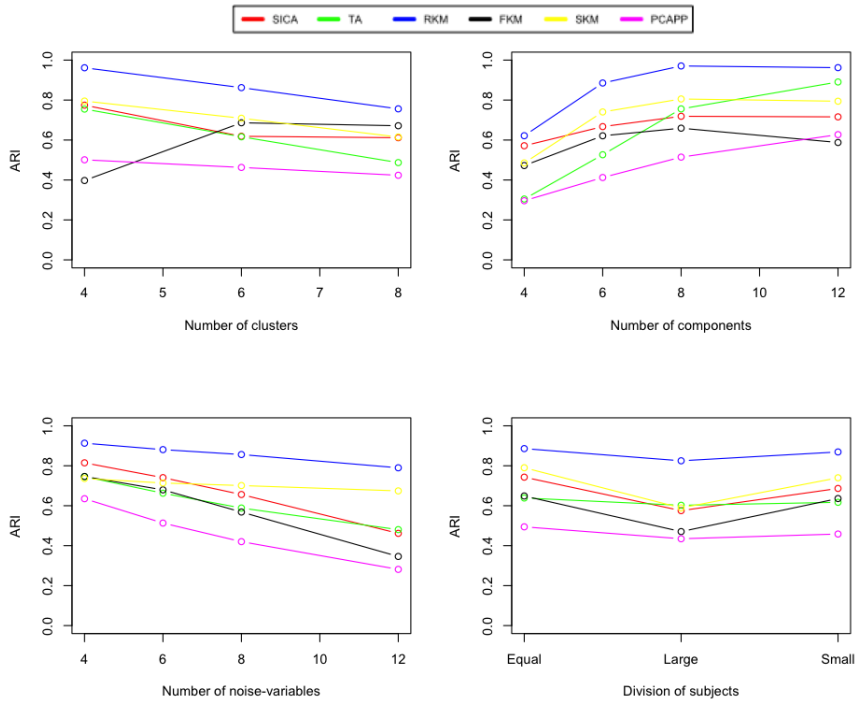

Figure 2. Average ARI values as a function of the method (lines in different colours) and number of clusters (upper left), number of components (upper right), number of noise variables (bottom left) and division of subjects across clusters (bottom right).

For the number of noise-variables present in the data, ANOVA demonstrates a substantial effect on goodness of recovery ($\eta^2_G = 0.44$, $F(1, 120) = 194{,}98$, $p < .001$). Table 1 in conjunction with the bottom left plot in Figure 2 shows that the goodness of recovery linearly deteriorates as the number of noise-variables increases for all methods but that the extent to which this decrease takes place differs per method ($\eta^2_G = 0.16$; $F(5, 600) = 44.45$, $p < .001$). SKM and RKM appear less sensitive to increases in noise depicted by relatively flat lines compared to the other methods that show more pronounced decreases in goodness

of recovery as noise increases. These results too, were found by Durieux and Wilderjans (2019) and Steinley et al. (2012).

The division of subjects per cluster, lastly, showed to be of main influence on the goodness of recovery since η^2_G= 0.17 ($F(2, 120) = 25.96$, $p < .001$), albeit to a lesser extent than the number of components or number noise variables, indicated by the smaller magnitude of the η^2_G coefficient. Regarding division of subjects over clusters, goodness of recovery was best when the division of subjects was equal over all clusters. The condition in which clusters consisted of one *small* cluster performed slightly worse and the condition in which one *large* cluster dominated the cluster subspace shows the worst cluster recovery. These results are visually represented in the bottom right panel in Figure 2, and confirm results found by Steinley et al. (2012) and are, most likely, due to the nature of the K-means algorithm itself and not inherent to the method of projection (Steinley et al., 2012). Note that no interaction was found between method and division of subjects.

The number of clusters present in the data showed no main effect, but did demonstrate an interaction effect with methods ($\eta^2_G = 0.31$; $F(5, 600) = 105.26$, $p < .001$). The interaction effect between the number of clusters present in the data and the method is presented visually in the upper left panel in Figure 2 using the data in Table 1. Similar to what Steinley et al. (2012) found, goodness of recovery decreases as the number of clusters present in the data increases. This is expected as the projection of points to a lower-dimensional space increases the chance of the clusters made up by these points overlapping, which is further supported by the fact that goodness of recovery increases as the number of variables making up the cluster space increases as shown in the upper right plot in Figure 2. In larger cluster-spaces, projections of clusters will, generally, show less overlap (Steinley et al., 2012). The upper left panel in

Figure 2 shows a linear decrease for all methods in terms of goodness of recovery except for FKM, which *increases* first, after which it flattens out.

Cluster validity Besides goodness of recovery of the true cluster structure, the properties of the obtained clusterings were evaluated using the SI (Rousseeuw, 1987). Table 2 shows the overall average Silhouette values for each of the methods and factor level.

Table 2

Average Silhouette Index values and standard deviations, overall and per method and factor level for the first simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.20	0.11	0.21	0.05	0.33	0.11	0.11	0.16	0.27	0.11	0.20	0.06	0.22	0.13
	6	0.22	0.09	0.20	0.04	0.35	0.09	0.23	0.19	0.26	0.10	0.20	0.06	0.24	0.12
	8	0.25	0.09	0.19	0.04	0.34	0.08	0.20	0.21	0.23	0.12	0.20	0.05	0.24	0.12
# Components	4	0.33	0.07	0.23	0.03	0.39	0.10	0.08	0.21	0.23	0.15	0.27	0.04	0.25	0.15
	6	0.25	0.07	0.20	0.04	0.39	0.07	0.23	0.20	0.31	0.10	0.21	0.04	0.27	0.12
	8	0.20	0.07	0.20	0.5	0.34	0.06	0.25	0.18	0.29	0.07	0.18	0.04	0.24	0.11
	12	0.12	0.03	0.17	0.03	0.23	0.03	0.14	0.11	0.19	0.04	0.14	0.03	0.17	0.07
# Noise variables	4	0.26	0.10	0.24	0.04	0.36	0.11	0.27	0.19	0.27	0.12	0.23	0.06	0.27	0.12
	6	0.24	0.10	0.21	0.03	0.35	0.10	0.23	0.19	0.25	0.12	0.21	0.05	0.25	0.12
	8	0.22	0.09	0.19	0.03	0.33	0.09	0.17	0.19	0.25	0.11	0.19	0.05	0.22	0.12
	12	0.18	0.08	0.17	0.04	0.32	0.08	0.06	0.14	0.25	0.10	0.17	0.05	0.19	0.12
Division	Equal	0.24	0.10	0.20	0.04	0.34	0.10	0.20	0.20	0.29	0.12	0.20	0.06	0.25	0.13
	Large	0.21	0.09	0.20	0.04	0.34	0.09	0.15	0.18	0.22	0.10	0.20	0.05	0.22	0.12
	Small	0.22	0.09	0.20	0.04	0.34	0.10	0.18	0.20	0.25	0.11	0.20	0.06	0.23	0.12
Overall		0.22	0.10	0.20	0.04	0.34	0.09	0.18	0.19	0.26	0.11	0.20	0.06	0.23	0.12

Note: Highest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

As was done for the goodness of recovery, ANOVA was conducted on the SI values and effects that are significant beyond the standard $\alpha = .05$ level and have a generalised eta-squared measure (η^2_G) larger than 0.15 will be discussed. ANOVA showed there were significant main- and interaction effects. Again, higher-order interactions were not deemed sufficiently interesting, and will not be discussed.

There were significant differences between methods in the validity of their obtained clusterings ($\eta^2_G = 0.42$; $F(5, 600) = 143.43$, $p < .001$). RKM showed

the highest overall cluster validity as its average SI value is 0.34, SKM follows with an average value of 0.26 and SICA is third in line with average overall SI values of 0.22. Note, this order of performance is identical to the one found for goodness of recovery.

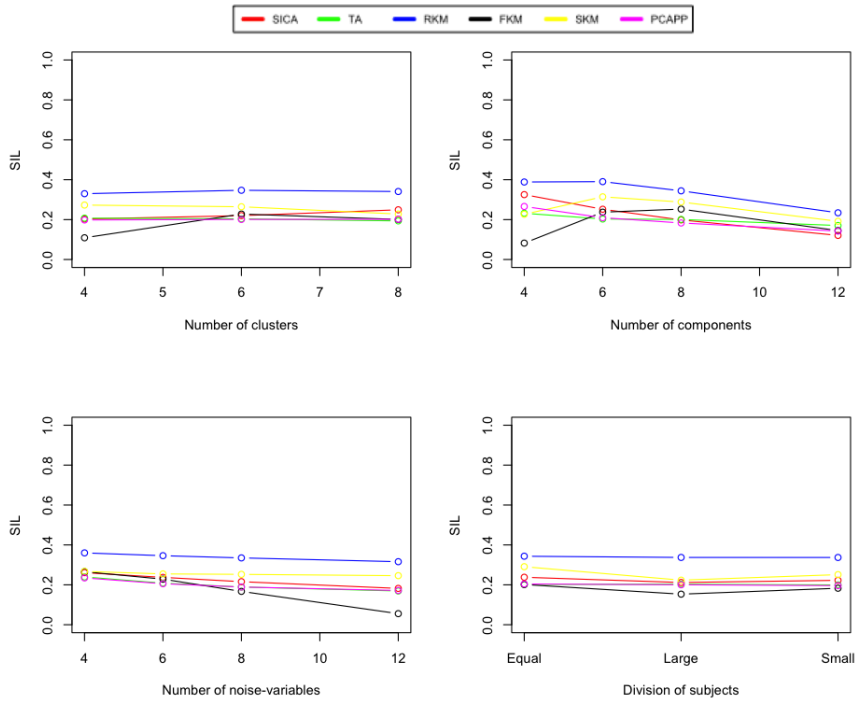

Figure 3. Average Silhouette values as a function of the method (lines in different colours) and number of clusters (upper left), number of components (upper right), number of noise variables (bottom left) and division of subjects across clusters (bottom right).

The number of components making up the cluster space showed a substantial main effect ($\eta^2_G = 0.24$, $F(1, 120) = 99.36$, $p < .001$) and a substantial interaction effect ($\eta^2_G = 0.17$ $F(5, 600) = 40.63$, $p < .001$). Table 2 together with the upper right panel in Figure 3 provide an interpretation to this effect. As can be

seen, all methods except FKM show an overall decreasing trend in cluster validity as the number of components making up the cluster space increases. The trajectory making up this trend, however, differs per method. PCAPP, SICA, TA and RKM show linear decreases, whereas SKM shows more of an inverted parabola. Silhouette values for SKM increase at first for 6 and 8 clusters, after which they decrease for 12 clusters. This pattern is very similar to that of FKM which demonstrates lowest Silhouette values for the four-component data, shows increased Silhouette values for 6 and 8 components, and shows a subsequential decrease at twelve components. Comparing this result to the influence of the number of components in the data on goodness of recovery, it is interesting to see that, except for FKM, cluster validity can *decrease* even when the goodness of recovery for those clusters *increases*. The fact that cluster validity decreases as dimensionality increases (that is, either by adding noise variables or multimodal components) could be due to the fact that cluster validity decreases in high dimensionality. This is what causes the curse of dimensionality described in the introduction.

The last main effect on cluster validity was found to be the number of noise variables present in the data ($\eta^2_G = 0.19$, $F(1, 120) = 72.37$, $p < .001$). Inspecting Table 2 and the bottom left panel in Figure 3, it is apparent that Silhouette values for all methods decrease as the number of noise variables present in the data increases. This trend is a linear one for all methods, but demonstrates to be especially pronounced for FKM. This is in line with the observation that goodness of recovery decreases as the number of noise variables increases.

No further main or interaction effects of practical significance were found in this simulation.

Bimodality of the extracted components Table 3 shows the average median BC values per method and factor level. As can be seen, SICA yields the

Table 3

Average median Bimodality Criterion values and standard deviations overall, and per method and factor level for the first simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.48	0.07	0.37	0.02	0.41	0.06	0.38	0.06	0.40	0.06	0.42	0.02	0.41	0.06
	6	0.49	0.06	0.37	0.02	0.41	0.05	0.44	0.07	0.41	0.05	0.43	0.02	0.42	0.06
	8	0.51	0.06	0.37	0.02	0.41	0.04	0.45	0.06	0.41	0.04	0.43	0.2	0.43	0.06
# Components	4	0.55	0.06	0.37	0.03	0.44	0.06	0.48	0.08	0.44	0.05	0.44	0.03	0.45	0.08
	6	0.51	0.06	0.37	0.03	0.42	0.04	0.43	0.06	0.43	0.04	0.43	0.02	0.43	0.06
	8	0.49	0.05	0.37	0.02	0.40	0.03	0.40	0.04	0.40	0.04	0.42	0.01	0.41	0.05
	12	0.44	0.05	0.36	0.02	0.37	0.02	0.37	0.02	0.37	0.02	0.41	0.01	0.39	0.04
# Noise variables	4	0.49	0.07	0.37	0.03	0.42	0.06	0.43	0.07	0.41	0.05	0.43	0.03	0.43	0.06
	6	0.49	0.07	0.37	0.02	0.41	0.05	0.43	0.07	0.41	0.05	0.43	0.02	0.42	0.06
	8	0.50	0.06	0.36	0.02	0.41	0.05	0.42	0.07	0.40	0.05	0.42	0.02	0.42	0.06
	12	0.51	0.05	0.36	0.02	0.40	0.04	0.40	0.06	0.40	0.04	0.42	0.01	0.42	0.06
Division	Equal	0.51	0.07	0.37	0.02	0.42	0.05	0.43	0.06	0.42	0.05	0.43	0.02	0.43	0.06
	Large	0.48	0.07	0.36	0.02	0.39	0.04	0.41	0.07	0.39	0.04	0.43	0.02	0.41	0.07
	Small	0.49	0.06	0.37	0.02	0.41	0.05	0.42	0.07	0.41	0.05	0.42	0.02	0.42	0.06
Overall		0.50	0.07	0.37	0.22	0.41	0.05	0.42	0.07	0.41	0.05	0.43	0.02	0.42	0.06

Note: Highest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

components demonstrating the highest BC values over all factors and levels. Note that, possibly due to taking the average median over the repetitions, no BC values in Table 3 for SICA exceed the 0.555 threshold set for bimodality (SAS, 2004). ANOVA main effects beyond the standard $\alpha = .05$ level and that have a generalised eta-squared measure (η^2_G) larger than 0.15 will be discussed. To save space, discussion of interaction effects for the multimodality of the extracted components has been added as Appendix B.

ANOVA showed that there was a significant and very substantial difference between methods in terms of their ability to extract multimodal components since ($\eta^2_G = 0.81$, $F(5, 600) = 857.66$, $p < .001$). This becomes apparent from Table 3 as the overall average median BC is highest for SICA, but does not differ substantially between the other methods. A main effect was also found for the number of components defining the cluster subspace ($\eta^2_G = 0.63$ $F(1, 120) = 510.61$, $p < .001$). As can be seen from the margin column in Table 3,

Table 4

Average median Dip test p-value and standard deviations, overall and per method and factor level for the first simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.49	0.28	0.89	0.09	0.75	0.19	0.81	0.20	0.77	0.19	0.87	0.12	0.76	0.23
	6	0.49	0.29	0.89	0.10	0.85	0.13	0.77	0.21	0.86	0.12	0.88	0.09	0.79	0.22
	8	0.54	0.33	0.89	0.10	0.85	0.14	0.78	0.21	0.87	0.12	0.88	0.10	0.80	0.22
# Components	4	0.31	0.27	0.84	0.14	0.70	0.21	0.60	0.27	0.73	0.21	0.83	0.15	0.67	0.28
	6	0.49	0.30	0.89	0.09	0.81	0.14	0.79	0.17	0.83	0.13	0.87	0.10	0.78	0.22
	8	0.52	0.28	0.90	0.07	0.86	0.11	0.86	0.11	0.87	0.10	0.87	0.08	0.82	0.20
	12	0.71	0.20	0.91	0.05	0.90	0.07	0.90	0.06	0.90	0.06	0.90	0.06	0.87	0.12
# Noise variables	4	0.57	0.31	0.88	0.11	0.80	0.18	0.76	0.23	0.82	0.17	0.85	0.13	0.78	0.22
	6	0.54	0.30	0.89	0.10	0.81	0.17	0.76	0.23	0.82	0.17	0.87	0.10	0.78	0.22
	8	0.50	0.29	0.88	0.09	0.83	0.15	0.79	0.19	0.84	0.14	0.88	0.10	0.79	0.22
	12	0.42	0.29	0.89	0.09	0.84	0.14	0.83	0.16	0.85	0.13	0.90	0.07	0.79	0.23
Division	Equal	0.38	0.29	0.88	0.10	0.80	0.18	0.77	0.22	0.80	0.18	0.88	0.10	0.75	0.26
	Large	0.66	0.24	0.88	0.10	0.85	0.14	0.84	0.16	0.86	0.14	0.87	0.10	0.83	0.17
	Small	0.48	0.30	0.89	0.09	0.80	0.16	0.75	0.23	0.84	0.13	0.87	0.11	0.77	0.23
Overall		0.51	0.30	0.89	0.10	0.82	0.16	0.79	0.21	0.83	0.15	0.87	0.10	0.78	0.22

Note: Lowest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

median BC values *decrease* as the number of components *increases*. The other factors showed no significant substantial main effects.

Much the same pattern appears in Table 4, showing median p-values averaged over the ten repetitions resulting from applying the Dip test to the components extracted by each of the methods. In agreement to the median BC values shown in Table 3, Dip test p-values are lowest for SICA and are substantially higher for all of the other methods. This reiterates the findings that SICA is able to identify non-Gaussian components to a larger extent than other methods, despite not always exceeding the necessary thresholds to formally conclude presence of non-Gaussian components (that is, exceed the 0.555 value set for BC values or the 0.05 level set for statistical tests for the Dip test).

Again, ANOVA revealed there was a very substantial difference between methods in terms of their average median p-values ($\eta^2_G = 0.80$, $F(5, 600) = 824.62$, $p < .001$). The number of components defining the subspace ($\eta^2_G =$

0.55; $F(1,\ 120) = 361.64,\ p < .001$) and the division of subjects over clusters ($\eta^2_G = 0.20;\ F(2,\ 120) = 36.84,\ p < .001$) also showed main effects. From the margin column of the table, it becomes apparent that, on average, the components extracted by all methods show more Gaussian shapes as the number of components defining the subspace increases, and that components show, on average, the most multimodality when the subjects are divided equally over the clusters, narrowly followed by the small condition, and worst for the large condition.

Multimodality and goodness of recovery A central notion in this thesis is that multimodal components are more interesting when aiming to find clusters in high dimensionality than unimodal/Gaussian components. To explore the relation between goodness of recovery and multimodality, correlations between BC values and ARI values, and between Dip test p-values and ARI values were calculated. Output showed a positive, highly significant Pearson correlation of r=0.09 between BC values and ARI values ($t(8638) = 8.03,\ p < .001$), meaning multimodality and goodness of recovery are positively associated, be it to a small extent. Conversely, there was a highly significant negative Pearson correlation of r= -0.07 between goodness of recovery and the p-values of the Dip test ($t(8638) = -6.89,\ p < .001$), demonstrating multimodal components are significantly but slightly associated with better goodness of recovery. It must be taken in consideration that, due to the large sample size in combination with the small magnitude of these coefficients, the practical implications this association is somewhat questionable. This is supported by the fact that SICA extracts components exhibiting the most multimodality (see Table 3 and Table 4) but that it does *not* outperform the other methods in terms of goodness of recovery (see Table 1).

Table 5

Average computation time and accompanying standard deviations, overall and per method and factor level for the first simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.26	0.09	0.15	0.03	15.98	4.45	18.07	3.12	6.17	1.66	2.53	1.29	7.20	7.64
	6	0.30	0.12	0.21	0.05	19.12	5.52	24.62	5.10	7.77	2.51	2.45	1.28	9.08	10.06
	8	0.30	0.11	0.24	0.04	28.05	16.25	40.43	25.35	9.46	2.96	2.48	1.30	13.49	19.75
# Components	4	0.19	0.04	0.16	0.03	31.59	17.11	41.28	30.01	9.53	3.74	1.52	0.67	14.04	21.67
	6	0.24	0.05	0.19	0.04	19.93	5.56	23.73	6.90	7.23	2.40	2.02	0.80	8.89	10.23
	8	0.30	0.07	0.21	0.05	16.88	5.09	22.49	5.83	6.95	1.75	2.56	0.93	8.23	9.16
	12	0.42	0.08	0.24	0.06	15.79	4.07	23.34	5.37	7.50	2.02	3.91	1.24	8.53	8.95
# Noise variables	4	0.24	0.09	0.19	0.05	19.07	16.57	27.56	20.54	7.46	2.72	1.61	0.74	9.36	15.05
	6	0.27	0.10	0.20	0.05	19.81	11.78	29.17	18.86	7.80	2.99	2.06	0.91	9.89	14.29
	8	0.29	0.10	0.20	0.06	20.99	3.40	28.99	16.76	7.84	2.81	2.54	0.98	10.14	13.49
	12	0.35	0.11	0.21	0.05	24.32	4.89	25.11	13.75	8.10	2.56	3.79	1.31	10.31	12.15
Division	Equal	0.29	0.11	0.20	0.05	21.00	10.78	28.45	18.87	7.03	2.67	2.55	1.34	9.92	14.12
	Large	0.29	0.10	0.19	0.05	20.67	11.79	26.21	13.67	8.79	2.81	2.48	1.26	9.77	12.65
	Small	0.29	0.10	0.21	0.06	21.47	11.71	28.46	19.92	5.58	2.56	2.47	1.28	10.08	14.53
Overall		0.29	0.11	0.20	0.05	21.05	11.43	27.71	17.72	7.80	2.78	2.50	1.29	9.92	13.79

Note: Lowest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

Computational efficiency The last evaluation metric used to measure performance of SICA and the other methods is computation time. As shown in Table 5, SICA converges to its final solution quickly due to the fastICA algorithm used in this thesis. However, TA is quicker due to the computational efficiency of PCA. All other methods take substantially more time to converge.

Evaluating performance of the first simulation Although RKM (and SKM to a smaller extent) clearly outperforms SICA in terms of cluster recovery and validity, it is more than 50 times (20 times for SKM) slower than SICA. TA is faster than SICA but recovers the clustering to a substantially lesser extent. Balancing computational efficiency and cluster recovery performance, SICA seems to be (one of) the best method. Posterior inspection of the generated data showed that not all of the variables that should contain cluster structure were multimodal, as depicted by the left boxplot in Figure 4 showing the distribution of BC-values of these variables. The majority of the BC-values

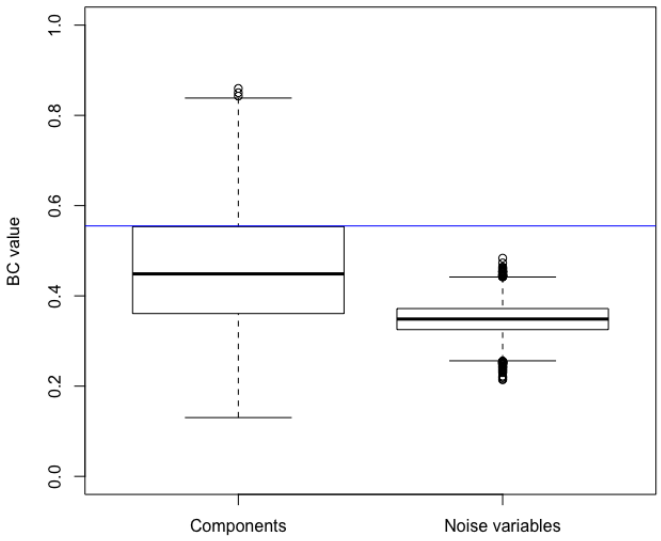

Figure 4. Boxplots of the BC-values for each component and noise variable generated by *clusterGeneration* for the 1440 data sets. The 0.555 threshold is depicted by the horizontal blue line and indicates that not all components defining the cluster subspace are multimodal.

are lower than 0.555, and all noise variables show BC-values lower than 0.555. In addition, the spread of BC-values is larger for the components than for the noise variables. SICA explicitly aims to find multimodal components (and does so better than other methods, as can be seen from Table 3 and Table 4), and the weaker performance of SICA in the first simulation could very well be caused by the fact that the components returned by *clusterGeneration* are not all non-Gaussian.

3.3 Simulation 2

As shown in Figure 4, the variables generated in the first simulation did not exhibit sufficient multimodality. This may have negatively impacted SICA performance in terms of goodness of recovery and cluster validity (see Table 1 and Table 2, respectively). In the second simulation, data will be generated using a procedure more adherent to the assumptions of SICA in terms of the multimodality of the components defining the subspace. It is expected that, by doing so, SICA will outperform the other methods in terms of goodness of recovery and cluster validity while attaining its computational efficiency and ability to extract multimodal components.

3.3.1 Data generation procedure

For the second simulation, data were generated based on the procedure outlined by Durieux and Wilderjans (2019) with some additional adaptations to fit the current factorial design. First, the binary partitioning matrix $\mathbf{P}$ ($N \times K$) was created based on the number of clusters and the division of subjects over these clusters with the constraint that, for each row of $\mathbf{P}$, the sum of the elements in each row equals 1 (i.e., that each of the subjects belonged to one and only one cluster). Next, a cluster centroid matrix $\mathbf{C}$ ($K \times Q$) was created where Q equals the number of components making up the space containing the cluster structure for each of the four levels (4, 6, 8 or 12 components) this factor takes on. Each row in $\mathbf{C}$ contains the centroid coordinates for one of the K clusters for which the total number is specified by the factorial design. All cluster centroid coordinates were sampled at random from the following set of numbers; [-20, -15, -10, -5, 0, 5, 10, 15, 20], meaning that a cluster centroid can, on any subspace component, have one of these values as its defining coordinate. No two rows in $\mathbf{C}$ are identical to make sure that cluster centroids do not perfectly

overlap (a property also found in the algorithm of Milligan, 1985). **P** and **C** are then multiplied and an error matrix **E** is added to this new matrix to make for the cluster space in which each of the clusters now contain some within-cluster variation. This error was generated by drawing independent numbers from $\mathcal{N}(0,1)$ and rescaled to match 5% within-cluster variance, denoting dense clusterings. Formally, the cluster structure **B** is then obtained as:

$$\mathbf{B} = \mathbf{PC} + \mathbf{E} \tag{14}$$

B was then evaluated with respect to which each of the components in **B** demonstrates sufficient cluster structure. Specifically, if the BC value for one of the components in **B** does not exceed 0.555, the centroid matrix **C** is iteratively sampled again and evaluated until every one of the components in **B** demonstrates sufficient cluster structure as indicated by the BC value. This implies that the clusters in a lower dimensional data space are generated in such a way that the variables have a multimodal distribution. An example of a two-dimensional cluster space containing four clusters generated by this procedure is given in Figure 5.

Next, the number of independent masking variables specified by the factorial design are generated as $\mathcal{N}(\mu, \sum)$ where $\mu = \mathbf{0}$ and $\sum = [Cov[M_i, M_j]$, with 100 for $i = j$ and 0 for $i \neq j]$ and concatenated to **B**. This combination is then multiplied by the square mixing matrix **A** generated by sampling independent numbers from U(-2,2). The dimensions of **A** were equal to ((number of components + number of noise variables) $\times$ (number of components + number of noise variables)), with the number of components and the number of noise variables equalling that of the level of the factorial design for which the data are gen-

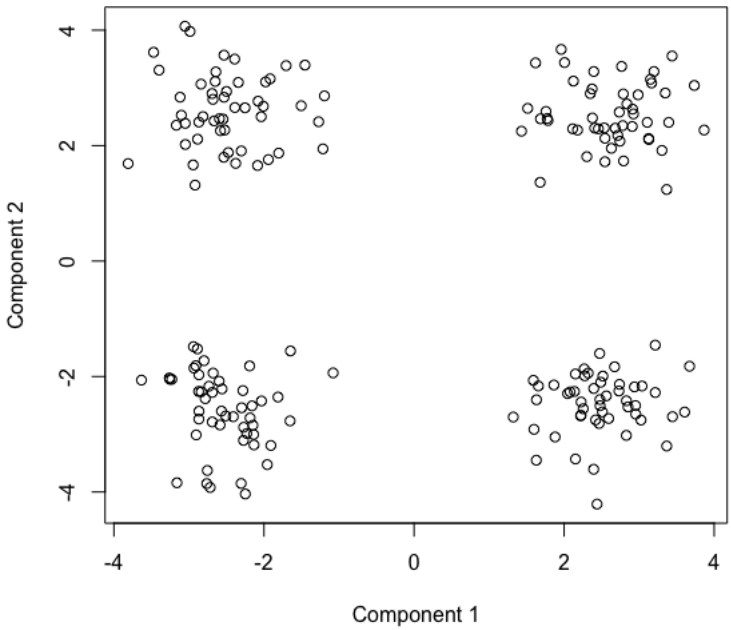

Figure 5. Example of a two-dimensional cluster space containing 200 subjects divided equally over four clusters generated by the procedure for simulation study 2.

erated. The eventual (high-dimensional) data matrix $\mathbf{X}$ was then constructed as:

$$\mathbf{X} = \mathbf{BA} \tag{15}$$

In which $\mathbf{B}$ now denotes the cluster space concatenated to the noise variables. As was done for the first simulation, the same combination of 1440 data sets

Table 6

Average ARI values and accompanying standard deviations, overall and per method and factor level for the second simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.90	0.16	0.78	0.23	0.80	0.22	0.12	0.16	0.56	0.22	0.74	0.26	0.65	0.33
	6	0.73	0.24	0.70	0.28	0.74	0.27	0.26	0.24	0.55	0.25	0.68	0.29	0.61	0.31
	8	0.72	0.27	0.65	0.28	0.69	0.27	0.30	0.28	0.51	0.24	0.63	0.28	0.58	0.30
# Components	4	0.60	0.26	0.39	0.18	0.44	0.19	0.16	0.15	0.28	0.15	0.38	0.19	0.38	0.24
	6	0.77	0.24	0.64	0.19	0.69	0.19	0.24	0.25	0.50	0.19	0.61	0.21	0.58	0.27
	8	0.86	0.18	0.82	0.16	0.85	0.15	0.26	0.25	0.64	0.18	0.77	0.18	0.70	0.28
	12	0.91	0.16	0.98	0.06	0.98	0.05	0.26	0.29	0.73	0.17	0.97	0.07	0.81	0.30
# Noise variables	4	0.85	0.21	0.77	0.24	0.79	0.23	0.40	0.30	0.58	0.23	0.77	0.23	0.69	0.29
	6	0.83	0.22	0.73	0.27	0.75	0.26	0.26	0.24	0.52	0.23	0.72	0.27	0.64	0.31
	8	0.78	0.24	0.69	0.27	0.72	0.26	0.18	0.18	0.53	0.24	0.67	0.27	0.59	0.32
	12	0.68	0.26	0.65	0.28	0.70	0.27	0.07	0.07	0.52	0.25	0.58	0.29	0.53	0.32
Division	Equal	0.89	0.19	0.76	0.25	0.80	0.23	0.24	0.25	0.61	0.24	0.73	0.27	0.67	0.32
	Large	0.63	0.26	0.64	0.28	0.67	0.27	0.20	0.20	0.44	0.20	0.63	0.28	0.53	0.30
	Small	0.84	0.19	0.73	0.27	0.76	0.26	0.25	0.27	0.46	0.25	0.70	0.28	0.64	0.32
Overall		0.78	0.24	0.71	0.27	0.74	0.27	0.23	0.24	0.54	0.24	0.68	0.28	0.61	0.32

Note: Highest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

were generated and subjected to SICA, TA, RKM, FKM, SKM and PCAPP using the true number of clusters and components. The number of subjects N was, for every data set, fixed at 200 and for all analyses, the number of within-cluster components for SKM was set at two. Evaluation of performance was conducted using the same metrics for goodness of recovery, cluster validity, multimodality of the components and computation time.

3.3.2 Results simulation 2

Goodness of recovery Table 6 shows the average ARI values and standard deviations, overall and per method and factor level. ANOVA on these results (only significant and with $\eta^2_G > 0.15$) revealed that a significant main effect of method was present ($\eta^2_G = 0.82$, $F(5, 600) = 1448.41$, $p < .001$). Inspection of the last row of Table 6 shows that SICA had the overall highest goodness of recovery as the overall average ARI value for SICA is 0.78. RKM (average

ARI of 0.74) follows shortly and TA (average of 0.71) and PCAPP (average of 0.68) at some more distance. FKM (average of 0.24) and SKM (average of 0.54) are clearly outperformed by SICA. No higher-order interactions were deemed sufficiently interesting (based on η^2_G higher than 0.30) to discuss.

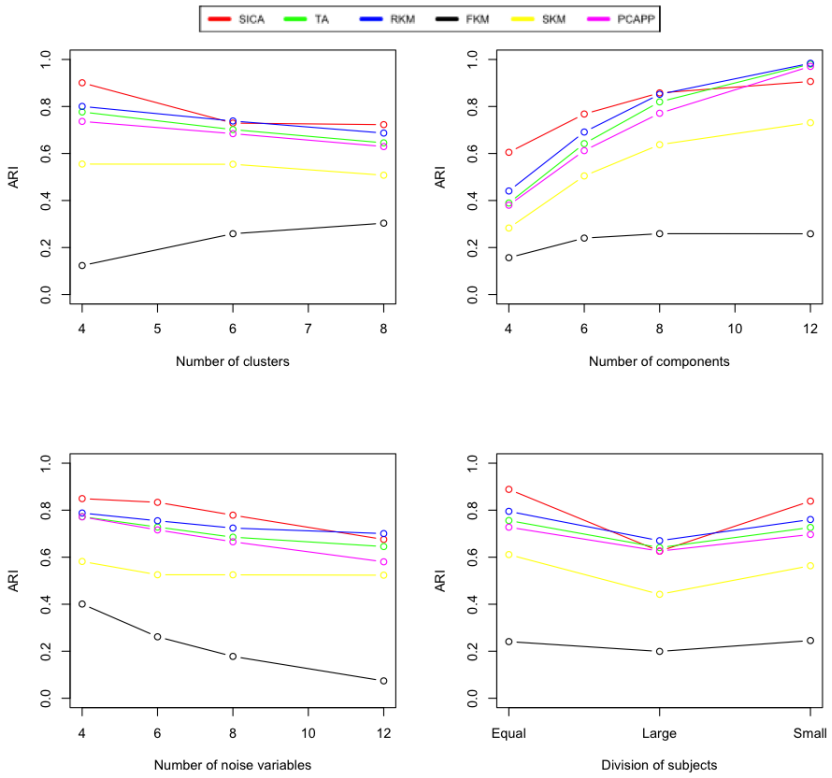

Figure 6. Average ARI values as a function of the method (lines in different colours) and number of clusters (upper left), number of components (upper right), number of noise variables (bottom left) and division of subjects across clusters (bottom right).

The effects of all factors on goodness of recovery, and their respective interactions with the factor method can be derived from Table 6 and Figure 6. The

number of components defining the subspace showed a main effect ($\eta^2_G = 0.75$; $F(1, 120) = 584.05$, $p < .001$) and interaction effect ($\eta^2_G = 0.35$, $F(5, 600) = 171.33$, $p < .001$). Average ARI values increase for all methods as the number of components increases (reiterating what was found in the first simulation), but the upper right panel in Figure 6 shows this to be especially pronounced for PCAPP, RKM, SKM and TA. FKM shows only a minor increase, as does SICA, for which the increase in average ARI values flattens out between 8 and 12 components.

The number of noise variables demonstrated a significant main effect ($\eta^2_G = 0.31$; $F(1, 120) = 86.02$, $p < .001$) and Table 6 shows average ARI values to decrease as the number of noise variables increases, as was the case in the first simulation. No substantial interaction was found between the factor method and the number of noise variables.

The division of subjects over the clusters showed a main effect on average goodness of recovery ($\eta^2_G = 0.31$, $F(2, 120) = 42.90$, $p < .001$) and Table 6 shows that all methods show the highest ARI values for equal division or small division and show notable decreases in goodness of recovery when the cluster structure is dominated by a single large cluster. Again, this confirms results from the first simulation and effects reported by Durieux and Wilderjans (2019) and Steinley et al. (2012). Division of subjects showed no substantial interaction effect with any of the other factors.

Lastly, despite not showing a substantial main effect ($\eta^2_G = 0.09$; $F(1, 120) = 18.60$, $p < .001$), the effect of the number of clusters differed significantly and substantially per method ($\eta^2_G = 0.23$, $F(5, 600) = 92.49$, $p < .001$) Inspecting the upper left panel in Figure 6 shows that, except FKM, which shows an increase in average ARI values as the number of clusters increases, all methods show a decrease in goodness of recovery for larger numbers of clusters. For TA, PCAPP

Table 7

Average Silhouette Index values and accompanying standard deviations, overall and per method and factor level for the second simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.25	0.11	0.32	0.09	0.40	0.09	-0.02	0.06	0.21	0.12	0.38	0.10	0.25	0.17
	6	0.27	0.10	0.30	0.09	0.36	0.09	0.03	0.15	0.20	0.13	0.35	0.09	0.25	0.16
	8	0.32	0.10	0.28	0.08	0.33	0.08	0.06	0.20	0.15	0.12	0.32	0.08	0.24	0.16
# Components	4	0.37	0.09	0.25	0.04	0.30	0.05	-0.07	0.06	0.07	0.08	0.31	0.06	0.20	0.17
	6	0.30	0.10	0.26	0.06	0.32	0.07	0.00	0.12	0.16	0.10	0.31	0.08	0.23	0.15
	8	0.25	0.08	0.29	0.07	0.37	0.08	0.06	0.14	0.22	0.10	0.34	0.09	0.26	0.14
	12	0.19	0.07	0.39	0.08	0.46	0.06	0.10	0.20	0.29	0.10	0.43	0.08	0.31	0.17
# Noise variables	4	0.31	0.11	0.36	0.09	0.41	0.10	0.13	0.20	0.20	0.15	0.42	0.09	0.30	0.17
	6	0.30	0.12	0.32	0.08	0.37	0.09	0.03	0.15	0.18	0.13	0.36	0.09	0.26	0.16
	8	0.27	0.10	0.28	0.07	0.35	0.08	-0.01	0.10	0.18	0.12	0.33	0.07	0.23	0.15
	12	0.23	0.09	0.24	0.05	0.33	0.07	-0.05	0.04	0.19	0.10	0.28	0.06	0.20	0.14
Division	Equal	0.32	0.12	0.31	0.08	0.38	0.09	0.02	0.15	0.22	0.13	0.36	0.10	0.27	0.17
	Large	0.24	0.08	0.29	0.09	0.35	0.09	0.01	0.14	0.15	0.11	0.33	0.09	0.23	0.15
	Small	0.28	0.11	0.30	0.08	0.36	0.09	0.03	0.17	0.19	0.12	0.35	0.09	0.25	0.16
Overall		0.28	0.11	0.30	0.08	0.36	0.09	0.02	0.15	0.19	0.13	0.35	0.09	0.25	0.16

Note: Highest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

and RKM, this decrease is linear, whereas for SICA it is more pronounced from 4 to 6 clusters, after which it flattens out for 6 and 8 clusters. These results reiterate those found in simulation 1 and those found by Steinley et al. (2012).

No higher-order interactions were deemed sufficiently interesting (based on η^2_G higher than 0.30) to discuss.

Cluster validity Cluster validity was evaluated using the SI (Rousseeuw, 1987). Table 7 shows the average Silhouette values overall, and per method and factor level.

ANOVA revealed multiple significant and substantial ($\eta^2_G > 0.15$) main- and interaction effects, no higher-order interactions were deemed sufficiently interesting (based on η^2_G higher than 0.30) to discuss. There was a substantial main effect of method on cluster validity ($\eta^2_G = 0.91$, $F(5, 600) = 2074.18$, $p <$ .001). This was to be expected given the range of average Silhouette values per method since RKM model output showed an average SI value of 0.36, whereas

this was only 0.02 for FKM and 0.19 for SKM.

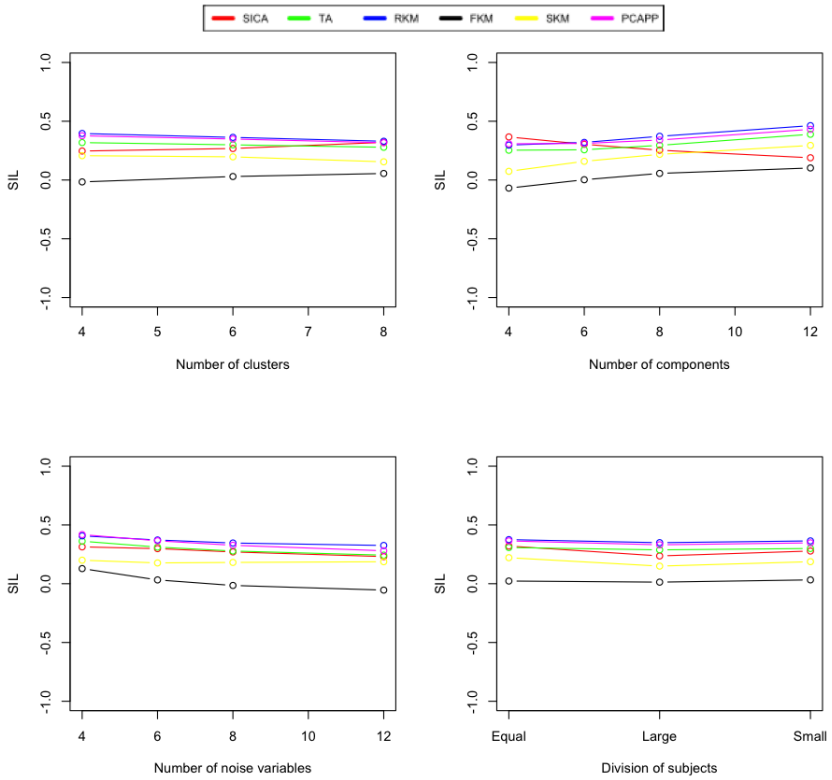

Figure 7. Average Silhouette values as a function of the method (lines in different colours) and number of clusters (upper left), number of components (upper right), number of noise variables (bottom left) and division of subjects across clusters (bottom right).

The division of subjects over the clusters also exercised a main effect over cluster validity ($\eta^2_G = 0.18$, $F(2, 120) = 33.37$, $p < .001$). The pattern observed in regard to this factor is similar to that observed for goodness of recovery. For equal and small divisions, cluster validity is similar but all methods except FKM show a notable decrease in cluster validity for the factor level in which one large

cluster dominates the division of subjects over clusters. This pattern was not observed this clearly in the first simulation.

The number of components also showed a main effect ($\eta^2_G = 56$; $F(1, 120) = 383.91$, $p < .001$) and an interaction effect ($\eta^2_G = 0.64$, $F(5, 600) = 349.10$, $p < .001$). From Table 7, it is apparent that average cluster validity increases as the number of components increases. The upper right panel in Figure 7, however, shows this to be the case for all methods except SICA, for which this pattern is the exact opposite. For SICA, average cluster validity decreases as the number of component increases, depicted by the linearly deteriorating red line. This result is contrary to what would be expected based on the Silhouette values in simulation 1. There, most methods showed an overall decrease in cluster validity as the number of components defining the subspace increased.

The number of noise variables, too, showed both a main ($\eta^2_G = 0.51$, $F(1, 120) = 318.19$, $p < .001$) and interaction ($\eta^2_G = 0.22$, $F(5, 600) = 54.90$, $p < .001$) effect. Average cluster validity deteriorates from 0.30 to 0.20 as the number of noise variables increases, as shown in Table 7. In addition, the bottom left panel in Figure 7 shows that all methods show an overall decrease in cluster validity as the number of noise variables increases. This is in accordance with earlier observations in this thesis that increasing noise in the data causes worse recovery of clusters and less clear clusterings. Especially for FKM, denoted by the black line, cluster validity quickly deteriorates as noise increases.

Despite showing no substantial main effect, an interaction was found between the methods and the number of clusters ($\eta^2_G = 0.32$, $F(5, 600) = 93.48$, $p < .001$). The upper left panel in Figure 7 shows that, except for SICA and SKM, all methods show linearly decreasing cluster validity as the number of clusters present in the data increases, confirming results found in simulation 1. FKM and SICA, instead, show linearly increasing cluster validity, which does not

Table 8

Average median Bimodality Criterion values and standard deviations overall, and per method and factor level for the second simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.51	0.08	0.37	0.02	0.38	0.03	0.38	0.05	0.38	0.03	0.50	0.10	0.42	0.08
	6	0.51	0.06	0.38	0.02	0.39	0.02	0.42	0.08	0.39	0.03	0.50	0.09	0.43	0.07
	8	0.57	0.07	0.38	0.02	0.39	0.03	0.42	0.07	0.39	0.03	0.41	0.08	0.44	0.08
# Components	4	0.57	0.08	0.38	0.03	0.39	0.03	0.40	0.06	0.40	0.04	0.48	0.05	0.43	0.08
	6	0.54	0.07	0.38	0.03	0.39	0.03	0.41	0.07	0.39	0.03	0.49	0.05	0.43	0.07
	8	0.52	0.06	0.37	0.02	0.39	0.02	0.41	0.08	0.38	0.03	0.49	0.05	0.42	0.08
	12	0.50	0.09	0.37	0.02	0.38	0.02	0.41	0.08	0.37	0.02	0.52	0.05	0.43	0.07
# Noise variables	4	0.52	0.09	0.38	0.03	0.39	0.03	0.45	0.09	0.39	0.04	0.53	0.05	0.45	0.09
	6	0.54	0.09	0.37	0.02	0.39	0.03	0.42	0.07	0.39	0.03	0.50	0.05	0.43	0.08
	8	0.53	0.07	0.37	0.02	0.38	0.02	0.39	0.05	0.38	0.03	0.49	0.04	0.42	0.08
	12	0.53	0.06	0.37	0.02	0.38	0.02	0.37	0.03	0.38	0.02	0.46	0.03	0.42	0.07
Division	Equal	0.57	0.08	0.38	0.02	0.39	0.03	0.40	0.07	0.39	0.03	0.50	0.05	0.44	0.09
	Large	0.50	0.07	0.37	0.02	0.38	0.02	0.40	0.08	0.38	0.03	0.49	0.05	0.42	0.07
	Small	0.53	0.07	0.37	0.02	0.39	0.03	0.41	0.07	0.39	0.03	0.49	0.05	0.43	0.08
Overall		0.53	0.08	0.37	0.02	0.39	0.03	0.41	0.07	0.39	0.30	0.49	0.05	0.43	0.08

Note: Highest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

match expectations based on the influence of numbers of clusters on goodness of recovery.

Bimodality of the extracted components Besides clustering quality, methods were also evaluated based on the extent to which the returned components show multimodal structure. Table 8 shows average median BC values averaged over methods and factor levels, and overall.

As can be seen in Table 8 by inspecting the underlined values, SICA extracts multimodal components to a larger extent than other methods, with two exceptions in which PCAPP yields components with higher BC values. The fact that PCAPP identifies multimodal components is expected; it maximises the CI which finds projections that deviate from Gaussian projections. All other methods show stable unimodality, returning homogeneous average median BC values and standard deviations.

ANOVA on these results revealed main- and interaction effects of the ma-

nipulated factors on the multimodality of the extracted components. To save space, discussion of all interaction effect has been added as Appendix C. Only the most important main effects will be discussed here.

As expected based on the results in Table 8, there was a significant difference in multimodality between the methods ($\eta^2_G = 0.91$, $F(5, 600) = 1508.37$, $p < .001$). This result is explained by the fact that PCAPP and SICA extract components that, on average, are more non-Gaussian than the components extracted by the other methods. It must be noted, however, that the BC values do not, for all cells in the design, exceed the 0.555 threshold set by SAS (2004) to define multimodality. This can be due to the fact that noise was added to the data, obscuring the multimodal structures present in the data or because the *average medians* are reported.

The number of noise variables present in the data also exercised a main effect ($\eta^2_G = 0.24$, $F(1, 120) = 163.91$, $p < .001$). Inspecting the margin column of Table 8, it appears that average median BC values somewhat deteriorate as the number of noise variables increase.

The second metric applied to quantify the extent to which the methods extract non-Gaussian subspaces is the Dip test (Hartigan & Hartigan, 1985). The average median p-values of the application of this test to the components extracted by each of the components are shown in Table 9.

As can be seen from Table 9, SICA components, on average, show higher multimodality than any other method as its overall average median p-value is 0.34, lower than any of the other overall average median p-values. ANOVA confirmed this difference to be very substantial ($\eta^2_G = 0.90$, $F(5, 600) = 1329.76$, $p < .001$).

As was the case for the BC-values, the number of noise variables exerts a substantial main effect on the extent to which extracted components were

Table 9

Average median Dip test p-value and standard deviations, overall and per method and factor level for the first simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
# Clusters	4	0.40	0.27	0.87	0.11	0.83	0.13	0.86	0.15	0.84	0.13	0,59	0.29	0.73	0.26
	6	0.36	0.27	0.88	0.09	0.85	0.11	0.75	0.27	0.85	0.11	0.64	0.25	0.72	0.27
	8	0.26	0.28	0.88	0.09	0.87	0.10	0.77	0.25	0.86	0.12	0.70	0.23	0.72	0.29
# Components	4	0.26	0.25	0.87	0.12	0.84	0.13	0.82	0.17	0.82	0.15	0.67	0.26	0.71	0.28
	6	0.29	0.27	0.87	0.10	0.85	0.12	0.79	0.24	0.85	0.13	0.65	0.26	0.72	0.28
	8	0.38	0.27	0.87	0.08	0.85	0.11	0.79	0.24	0.86	0.10	0.66	0.25	0.73	0.26
	12	0.44	0.28	0.88	0.08	0.86	0.09	0.78	0.26	0.86	0.09	0.58	0.26	0.73	0.26
# Noise variables	4	0.40	0.31	0.86	0.11	0.83	0.13	0.67	0.30	0.82	0.14	0.46	0.30	0.67	0.29
	6	0.35	0.28	0.87	0.11	0.85	0.11	0.78	0.24	0.86	0.10	0.63	0.25	0.72	0.27
	8	0.34	0.26	0.88	0.09	0.85	0.10	0.85	0.17	0.84	0.12	0.69	0.21	0.74	0.26
	12	0.27	0.24	0.98	0.09	0.87	0.10	0.87	0.10	0.86	0.10	0.78	0.16	0.76	0.26
Division	Equal	0.21	0.24	0.87	0.10	0.83	0.12	0.77	0.25	0.83	0.13	0.61	0.27	0.69	0.30
	Large	0.51	0.24	0.89	0.09	0.87	0.10	0.83	0.19	0.86	0.11	0.68	0.25	0.77	0.22
	Small	0.30	0.26	0.87	0.11	0.84	0.12	0.78	0.25	0.85	0.11	0.64	0.27	0.71	0.28
Overall		0.34	0.28	0.88	0.10	0.85	0.12	0.80	0.23	0.85	0.12	0.64	0.26	0.72	0.27

Note: Lowest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

multimodal ($\eta^2_G = 0.18$; $F(1, 120) = 137.62$, $p < .001$), with average median p-values *increasing* as the number of noise variables increases, meaning that the components become more unimodal.

A main effect of the division of subjects over clusters was also found ($\eta^2_G = 0.25$; $F(2, 120) = 100.25$, $p < .001$) and the margin column in Table 9 shows that the equal condition yielded the lowest average median p-values, followed by the small condition and worst for the large condition.

Multimodality and goodness of recovery As was done in the first simulation, the association between multimodality of the extracted components and goodness of recovery for each of the methods was explored using correlation analysis. The correlation between BC values and ARI values was highly significant and positive (r=0.32) as $t(8638) = 31.40$, $p < .001$, indicating a medium association between multimodality and goodness of recovery. Association was also observed for median Dip test p-values and ARI values as r = -0.32 ($t(8638)$

Table 10

Average computation time and accompanying standard deviations, overall and per method and factor level for the second simulation study.

Factor	Levels	SICA		TA		RKM		FKM		SKM		PCAPP		Overall	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
# Clusters	4	0.25	0.09	0.11	0.02	15.49	10.39	15.88	3.40	9.71	2.39	2.75	1.59	7.37	8.13
	6	0.28	0.09	0.15	0.02	28.82	19.84	31.52	17.71	10.84	2.46	2.75	1.57	12.39	17.04
	8	0.26	0.08	0.18	0.03	72.38	50.70	73.28	52.23	11.78	2.75	2.71	1.53	26.77	44.27
# Components	4	0.19	0.06	0.15	0.04	48.40	44.27	68.63	63.15	11.87	2.93	1.55	0.68	21.80	41.41
	6	0.24	0.06	0.15	0.04	36.81	45.39	39.54	29.19	10.58	2.56	2.12	0.93	14.91	27.73
	8	0.28	0.06	0.15	0.04	29.93	29.10	28.52	16.86	10.08	2.33	2.82	1.08	11.96	18.69
	12	0.35	0.07	0.14	0.04	40.46	37.70	24.22	10.01	10.57	2.53	4.46	1.58	13.37	21.64
# Noise variables	4	0.21	0.07	0.14	0.04	58.46	53.14	44.11	42.49	10.45	2.56	1.67	0.85	19.18	36.28
	6	0.24	0.08	0.15	0.04	38.23	36.62	44.20	38.10	10.81	2.74	2.23	1.03	15.31	27.66
	8	0.27	0.07	0.15	0.04	33.73	34.70	40.19	42.84	10.69	2.62	2.80	1.24	14.64	27.78
	12	0.34	0.08	0.16	0.04	25.17	22.30	36.40	36.13	11.16	2.74	4.23	1.70	12.91	22.03
Division	Equal	0.25	0.09	0.15	0.04	35.12	33.66	39.70	39.70	9.90	2.50	2.69	1.59	14.63	29.91
	Large	0.28	0.08	0.15	0.04	46.99	48.69	38.35	35.27	11.81	2.65	2.77	1.57	16.56	30.84
	Small	0.26	0.08	0.15	0.04	35.59	35.59	42.63	44.59	10.62	2.53	2.76	1.53	15.34	29.02
Overall		0.26	0.09	0.15	0.04	38.90	40.16	40.23	40.05	10.78	2.68	2.74	1.56	15.51	28.97

Note: Lowest value per factor level is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit.*

= -31.41, $p < .001$), indicating a medium effect too.

Computational efficiency The last metric used to evaluate model performance is the computational efficiency of each of the models. Table 10 shows the average computation time and standard deviations overall and per method and factor level. The pattern observed is the same as that of simulation 1, TA and SICA demonstrate low computational taxation whereas all other methods show significantly longer average computation times (about 10 times more for PCAPP, 40 times more for SKM and more than 100 times for RKM and FKM). Balancing cluster recovery and computational efficiency, SICA is clearly the best performing method as it is substantially faster than the other methods (except TA) and (clearly) outperforms all the other methods (see Table 6).

Evaluating performance of the second simulation SICA performance in terms of goodness of recovery was superior in the second simulation but not in the first. The procedure of data generation applied in the second simulation

adhered more closely to the assumptions of SICA in terms of the multimodality of the generated independent components. As a consequence, in simulation two, SICA outperformed traditional methods both in terms of goodness of recovery and in terms of computational efficiency. In the first simulation, only the latter was the case. Based on the combination of results from the two simulations, it is evaluated that SICA is an attractive alternative to traditional methods in that sense that, in the case of Gaussian subspaces (simulation 1), it challenges performance of the other methods, and in the case of non-Gaussian subspaces (simulation 2), it outperforms the other models while also retaining its computational efficiency.

4 Analysis of empirical data

In this section, the aforementioned models will be applied to three psychological data sets to compare performance of SICA to other methods. As will be discussed, for the fMRI applications, the number of clusters is set to the number of diagnoses present in the data but the number of components is decided based on model selection. Vice versa, for the sensory-processing sensitivity application, the number of components is set, whereas the number of clusters will be based on model selection.

4.1 Application to two-group fMRI data

In this section, SICA will be applied to fMRI data containing diagnoses on Alzheimers' Disease (AD) and elderly healthy control (HC) patients from the Medical University of Graz, Austria. These data were collected as part of the the prospective registry on dementia (PRODEM; Seiler et al., 2012) project and contain multiple modalities of MRI for 77 clinically diagnosed probable AD patients and 173 normal elderly controls (N=189) and were pre-processed by de Vos et al. (2016). Data were only used for AD patients for which diagnoses were conducted on the basis of the NINCDS-ADRDA criteria (McKhann et al., 2011). Grey matter density (GMD) values were used since grey matter atrophy is one of the main characteristics of AD and has been found to discriminate AD well from healthy control subjects (Cuingnet et al., 2011). For each subject, voxel-wise averaged GMD values were computed for each of the 48 regions in the probabilistic Harvard-Oxford cortical atlas yielding 48 GMD values per subject (de Vos et al., 2020).

Data were subjected to K-means on the full data, SICA, TA, RKM, FKM, SKM and PCAPP with the number of clusters set to 2, one for each of the clinical diagnoses present in the data. For each of the models, two components

were selected because, as shown in Appendix D.1, model performance did not increase substantially for larger number of components, and because using two components allows for easy visual interpretation. For SICA and PCAPP, components were selected based on the highest BC and CI values, respectively, and the number of within-cluster components for SKM was, for all combination of clusters and components, set to 2. Model performance was evaluated based on the ARI-values (Hubert & Arabie, 1985) of the obtained clusters compared to the diagnoses and the Silhouette indices of the obtained clusters (Rousseeuw, 1987). The components extracted by each of the models were compared based on their BC-value (Freeman & Dale, 2013; SAS, 2004) and the p-values resulting from the Dip test (Hartigan & Hartigan, 1985).

Table 11

ARI, SI, BC and Dip test p-values for the two extracted dimensions for each method in the two-group problem.

Method	ARI	Silhouette	BC dim. 1	BC dim. 2	Dip p-value dim. 1	Dip p-value dim. 1
SICA	<0.01	0.37	0.53	0.47	0.36	0.80
PCAPP	0.01	0.40	0.40	0.40	0.15	0.95
TA	0.39	0.43	0.39	0.35	0.94	0.99
RKM	0.41	0.53	0.40	0.35	0.93	0.38
FKM	0.00	0.44	0.46	0.34	<.001	0.98
SKM	0.04	0.16	0.34	0.30	0.97	0.84
KM	0.41	0.16				

Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit, KM = K-means.

Table 11 shows the ARI-values, Silhouette-values, BC-values and Dip test p-values for each of the models for the two-cluster problem. ARI-values compare the true diagnostic labels with the obtained clustering and shows that RKM and K-means yield the highest agreement between the actual diagnoses and the predicted clusterings. SICA yields (very) low agreement between these two, showing ARI-values close to chance (close to 0). Silhouette values are the high-

est for RKM and SICA performs in the middle when comparing the methods in terms of cluster validity as the Silhouette value is 0.37. SICA does, as expected, yield the highest BC-values for its extracted components, demonstrating the most multimodal components. These values do not, however, exceed the 0.555 threshold (SAS, 2004). Lastly, Dip test p-values are lowest for RKM and FKM. Only the first FKM component, however, is lower than the common .05 threshold set for p-values and so only this component demonstrates multimodality. It should be noted that this is unexpected since FKM also capitalises on PCA and does not explicitly aim to find non-Gaussian subspaces (as SICA is doing).

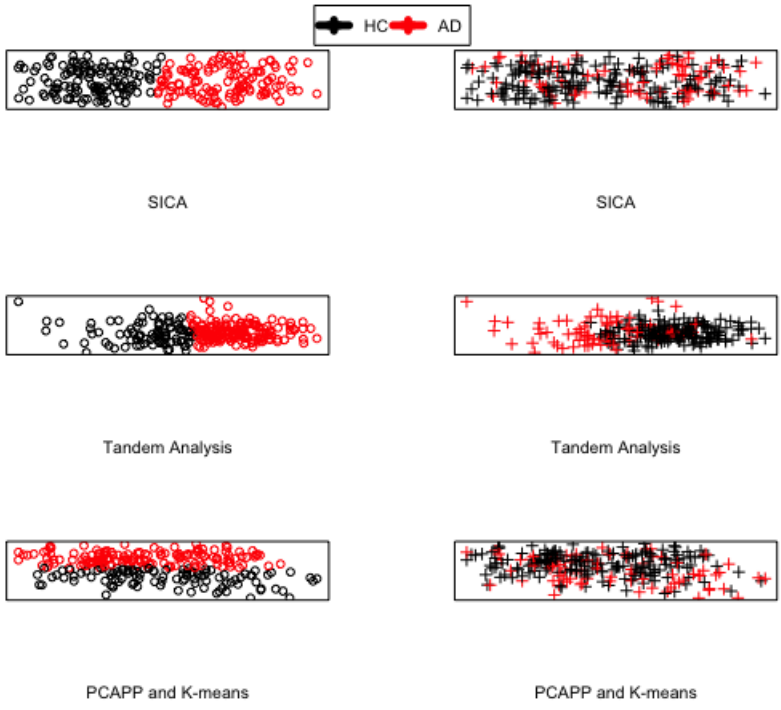

Figure 8. Results of applying SICA, TA and PCAPP to the two-cluster Graz data. Obtained clusterings are shown on the left and true diagnoses shown on the right, with the axes being the components extracted by each method.

To study why SICA and PCAPP fail to discover the two-cluster structure, Figure 8 shows the true diagnoses on the right and the obtained clusterings on the left for the two dimensions obtained by SICA, PCAPP and TA (to save space, figures for RKM, FKM, SKM and KM have been moved to Appendix E). For SICA and PCAPP, no clear cluster structure can be detected but the cluster structure identified by the PCA (in the first step of TA) is recognised by TA quite well. This reiterates the relatively large ARI value for TA in the two-cluster problem visible in Table 11. The high ARI values for KM and RKM are reiterated in the agreement between the actual and the obtained clusterings by these methods visible from Figure E.1 and Figure E.2 in Appendix E. Conversely, low ARI values for SKM and FKM are depicted in Figure E.1 and Figure E.2 as showing little agreement. It can be concluded that it seems that the components of the two-dimensional subspace in which the two diagnostic clusters reside have a more or less Gaussian shape and, therefore, cannot be identified by SICA and PCAPP.

4.2 Application to three-group fMRI-data

In this section, SICA will be applied to fMRI data containing diagnoses on AD as well. However, unlike the data used in the previous section, the clinical diagnoses present in these data are AD, mild cognitive impairment (MCI) and subjective memory complainers (SMC). Data were used from the Leiden-Alzheimer Research Nederland (LeARN) project (Handels et al., 2012; Jansen et al., 2017). These data contain multiple modalities of MRI for 61 probable Alzheimers' Disease (AD) patients, 61 patients with mild cognitive impairment (MCI) and 67 subjective memory complaining (SMC) subjects ($N = 189$). AD diagnosis was conducted based on the NINCDS-ADRDA criteria (McKhann et al., 2011), SMC diagnosis was done based on the clinical criteria for AD-caused

MCI (Albert et al., 2011), and subjects that did not meet criteria for either AD or MCI were included in the SMC group. Data were preprocessed by de Vos et al. (2020). Similar to the two-cluster problem, GMD values were used and evaluation of model performance is conducted using the same metrics as used in the previous section. All models were applied using three clusters (based on the presence of clinical diagnoses as described) and two components to aid in interpretation and because, as shown in Appendix D.2, model performance did not increase substantially for larger number of components.

Table 12

ARI, SI, BC and Dip test p-values for the two extracted dimensions for each method in the three-group problem.

Method	ARI	Silhouette	*BC* dim. 1	*BC* dim. 2	Dip *p*-value dim. 1	Dip *p*-value dim. 1
SICA	0.01	0.34	0.53	0.24	0.23	0.86
PCAPP	0.01	0.44	0.40	0.42	0.99	0.92
TA	0.06	0.36	0.31	0.31	0.82	0.99
RKM	0.08	0.42	0.32	0.31	0.89	0.99
FKM	-0.01	-0.03	0.50	0.48	0.07	0.69
SKM	0.01	0.11	0.31	0.32	0.99	0.43
KM	0.08	0.09				

Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit, KM = K-means.

Table 12 shows the ARI-values, Silhouette-values, BC-values and Dip test p-values for each of the models for the three-cluster problem. ARI-values compare the true diagnostic labels with the obtained clustering and shows that RKM and K-means yield the highest agreement between the actual and the predicted clusterings, although agreement is very low (ARI below .10) in general. SICA yields very low agreement between these two, showing ARI-values close to chance (close to 0). Silhouette values are the highest for PCAPP for the three cluster-problem and SICA does outperform some (but not all) methods, as its Silhouette value is 0.34.

SICA does, as expected, yield the high BC-values for its extracted components. However, the second dimension of FKM yields higher BC values than the second dimension for SICA. Note that none of the extracted components exceeds the 0.555 threshold (SAS, 2004). Lastly, Dip test p-values are lowest for FKM and SKM for the three-cluster problem, which is an unexpected result.

To further investigate the weak performance of SICA, Figure 9 shows the true diagnoses on the right and the obtained clusterings for each subject on the left on the two dimensions obtained by SICA, PCAPP and TA for the three-cluster problem (to save space, figures for RKM, FKM, SKM and KM have been moved to Appendix F). For all three methods shown in Figure 9, it is true that the retained lower-dimensional subspace does almost contain no cluster structure, which explains the very weak cluster recovery performance of the methods. This is equally the case for RKM, FKM, SKM and KM (visible in Figure F.1 and Figure F.2 in Appendix F, respectively). The visual representations of the actual versus the obtained clusterings show little agreement, explaining the ARI values at chance level (or only slightly above, in the case of RKM and KM).

4.3 Sensory-processing sensitivity

In this section, SICA performance will be compared to performance of the other models discussed so far on sensory-processing sensitivity (SPS) data. In this illustrative example, however, no diagnoses or 'true' clusters are present with which obtained clusterings can be compared, making it more akin to a realistic academic or applied scenario in which the number of clusters has to be decided based on model fit. As will be described, the number of components for all methods will, however, be set. Moreover, cluster recovery cannot be used as a performance metric and cluster validity remains as the sole performance criterion.

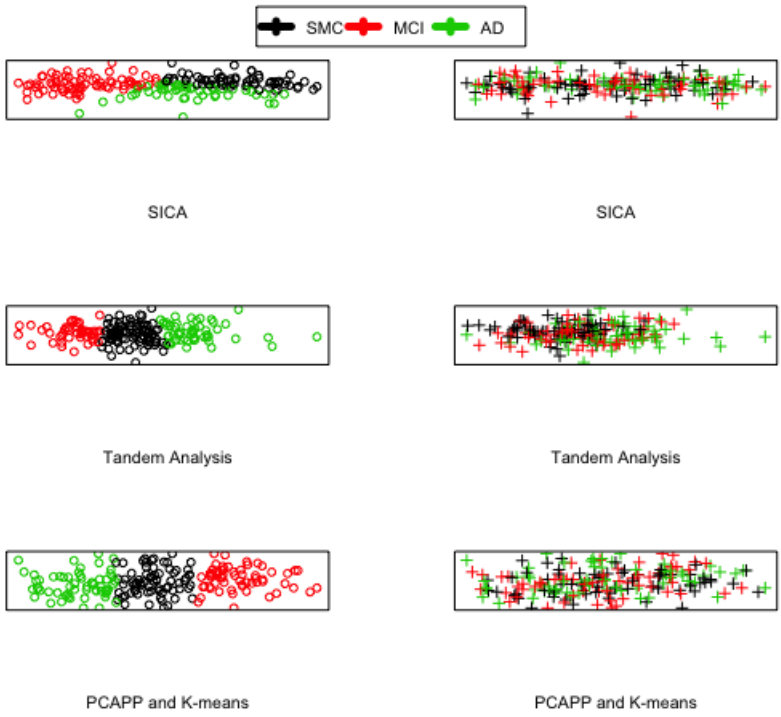

Figure 9. Results of applying SICA, TA and PCAPP to the three-cluster LeARN data. Obtained clusterings are shown on the left and true diagnoses shown on the right, with the axes being the components extracted by each method.

The data consist of 998 subjects measured on the Highly Sensitive Person Scale (HSPS; Aron & Aron, 1997) which is a scale measuring SPS using 27 items rated from 0 to 7 (”Strongly disagree” to ”Strongly agree”; Smolewska, McCabe, & Woody, 2006). Psychometric inspection of the HSPS using factor analysis shows it can be decomposed into three underlying factors/components with differing substantive meaning in terms of SPS and relation to personality traits (Smolewska et al., 2006). Therefore, the number of components for each of the methods applied in this section was set at three.

All models were applied to the data with a combination of three components and a number of clusters varying from two to six, where model-performance was evaluated using the Calinski-Harabasz criterion (Caliński & Harabasz, 1974). For each of the models, the number of clusters was selected that yielded the highest Calinski-Harabasz criterion, subsequently subjected to suitable methods of evaluation. For a full comparison of model performance per number of clusters, we refer the reader to Appendix D.3. Obtained clusterings were compared using the SI and, for each of the three extracted components, their BC-value and Dip test p-value was computed.

Table 13

BC values and Dip test p-values for each of the three extracted dimensions for each method.

Method	BC Dim. 1	BC Dim. 2	BC Dim. 3	p-value Dim. 1	p-value Dim. 2	p-value Dim. 3
SICA	0.58	0.39	0.33	<.001	0.96	0.99
PCAPP	0.42	0.39	0.41	0.96	0.99	0.99
TA	0.42	0.30	0.30	0.99	0.99	0.98
RKM	0.42	0.30	0.30	0.99	0.39	0.99
FKM	0.31	0.29	0.27	0.95	0.87	0.74
SKM	0.45	0.36	0.28	0.96	0.98	0.99

Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit, KM = K-means.

Table 13 shows the BC-values and Dip test p-values for each of the three extracted components for each method. The first SICA dimension exceeds the 0.555 threshold, demonstrating bimodality (SAS, 2004), a result which is confirmed by the p-value returned by the Dip test, illustrating the first component is at least bimodal (Hartigan & Hartigan, 1985). For the second and third dimensions, no method extracts dimensions that exceed the bimodality threshold or the 0.05 significance-threshold for the Dip test. SICA does, together with PCAPP, yield the most bimodal second dimension but this is not sufficient.

Similarly, the third component extracted by SICA has a lower BC value than that of PCAPP. All PCA-based methods extract components that do not exceed either the BC or Dip test threshold for bimodality/multimodality. This was expected, since PCA extracts predominantly Gaussian, unimodal components.

Table 14

Silhouette index values and the number of clusters selected for each model.

Method	Silhouette	Number of clusters
SICA	0.27	6
PCAPP	0.33	5
TA	0.48	2
RKM	<u>0.51</u>	2
FKM	0.25	5
SKM	0.09	2
KM	0.23	2

Note: Highest value is underlined. *Abbreviations: SICA = Subspace Independent Component Analysis, TA = Tandem Analysis, RKM = Reduced K-means, FKM = Factorial K-means, SKM = Subspace K-means, PCAPP = Principal Cluster Axis Projection Pursuit, KM = K-means.*

Cluster validity was assessed using the SI (Rousseeuw, 1987) and is shown in Table 14. As can be seen, RKM yielded the highest Silhouette value (0.51), whereas SICA yields a Silhouette value of 0.27. PCAPP and TA, in turn, outperform SICA as well, yielding Silhouette values of 0.33 and 0.48, respectively. For more detailed interpretation, Silhouette plots have been added in Appendix G. Inspecting Table 14, it is interesting to see that TA, RKM, SKM and KM deliver solutions with two clusters, whereas PCAPP, SICA and FKM deliver more complex solutions with up to six clusters.

5 Discussion

In this thesis, SICA, a novel method for finding clusters in non-Gaussian subspaces for high dimensional data, is proposed. SICA uses the Bimodality Coefficient (BC) to extract multimodal independent components that may contain relevant clustering information and performs K-means on the selected components. The merits of SICA have been thoroughly tested in two simulation studies and applied to three empirical settings using data on Alzheimer's Disease and Sensory Processing Sensitivity. In these studies, SICA is also compared to traditional methods that combine dimension reduction with clustering.

SICA performance was evaluated in comparison to performance of other methods aimed at finding clusterings in lower-dimensional projections of the original data based on *(1)* goodness of cluster-recovery, *(2)* the properties of the obtained clusters, *(3)* the extent to which the extracted components were of non-Gaussian shape, and *(4)* computational efficiency of the algorithms. Simulation studies allowed to systematically compare these metrics in a wide range of scenarios regarding *(1)* the number clusters in the data, *(2)* the number of components defining the cluster subspace, *(3)* the amount of error present in the data and *(4)* the division of subjects over the clusters. Empirical applications allowed to evaluate SICA in scenarios where the number of clusters or components have to be defined based on model selection, and illustrated its usefulness in applied or academic settings.

Results show that SICA is well able to recover obscured cluster structures in high-dimensional settings but that the extent to which it is capable of doing so depends on the data model and the characteristics of that data defined in the study design.

Goodness of recovery was specifically investigated in regard to increasing numbers of clusters in the data and increasing numbers of noise variables. Re-

garding the number of noise variables present in the data, all methods demonstrated a decrease in goodness of cluster recovery as more noise variables were present in the data, including SICA. This pattern was expected as it was reported earlier by Durieux and Wilderjans (2019) and Steinley et al. (2012). For SICA, this was especially pronounced in simulation 1, but less in simulation 2. Regarding the number of clusters present in the data, all methods showed decreased goodness of recovery as the number of clusters in the data increased (FKM being an exception in simulation 1), confirming, again, what was observed by Durieux and Wilderjans (2019) and Steinley et al. (2012). SICA suffered less from this influence in simulation 2 than in simulation 1. Results from both simulations allow to conclude that the hypotheses regarding goodness of recovery in section 3.2 seem to hold; goodness of recovery decreased for all methods as the number of noise variables increased and as the number of clusters present in the data decreased. For the other manipulated factors, no formal hypothesis were formulated but these demonstrated the same effect in both simulations; goodness of recovery increased as the number of components defining the subspace increased and was the best for the equal division of subjects over the clusters and the worst for the condition in which one large cluster dominated the cluster subspace. For the empirical fMRI data, SICA did not show high goodness of recovery, contrary to the simulation studies. Application to empirical data naturally carries the risk that data contain no (underlying) cluster structure, making application of cluster algorithms of little use. It must be acknowledged, however, that other methods such as RKM did manage to find underlying cluster structures. A possible explanation for this finding is that the components of the subspace in which the clusters reside are Gaussian, which prevents SICA from correctly retrieving these components and the associated cluster structure.

Besides goodness of recovery, the validity of the obtained clusterings was investigated for each of the methods using the Silhouette Index (Rousseeuw, 1987). SICA did not demonstrate superiority in terms of validity of its obtained clusterings as RKM mostly obtained more compact clusterings. In regard to the manipulated factors, cluster validity decreased as the number of noise variables increased but showed little sensitivity to the number of clusters present in the data, but did demonstrate an interaction effect in simulation 2, differing significantly per method and number of clusters. The influence of the number of components defining the cluster subspace revealed incoherent as increasing numbers of components made for *decreasing* cluster validity in simulation 1 but *increasing* cluster validity in simulation 2. Division of subjects over clusters showed the same effect on cluster validity as it did for all performance metrics as average Silhouette values decreased for the condition in which the cluster subspace was dominated by a single cluster. Application to sensory processing-sensitivity and Alzheimer's Disease data, however, showed SICA did not yield high Silhouette values.

The extent to which each method was able to identify a low-dimensional representation containing cluster structure was also investigated using the *Bimodality Coefficient* (Freeman & Dale, 2013; SAS, 2004) and the Dip test (Hartigan & Hartigan, 1985). Results were clear in this regard. SICA extracts components demonstrating the highest multimodality (and non-Gaussianity) of all other methods in both simulations, confirming the hypothesis. The influence of the number of noise variables and clusters on the multimodality of the extracted components was less clear. Both simulations showed that the multimodality of the components extracted by SICA increased as the number of clusters increased, confirming expectations. However, as the number of noise variables present in the data increased, simulation 1 showed *decreasing* multimodality of

the extracted components, whereas simulation 2 showed *increasing* multimodality. The hypothesis regarding the influence of the number of clusters present in the data on multimodality can, therefore, be confirmed, whereas the hypothesis concerning the number of noise variables cannot be considered confirmed. Application to empirical data shows SICA identifies multimodal components well, although not all components exceeded the thresholds for multimodality.

Concluding, results show SICA is a computationally fast algorithm for identifying non-Gaussian subspaces, and that subsequent application of K-means can identify clusterings to a large extent, especially when the true components making up the subspace are non-Gaussian (and multimodal). This combination does not always outperform similar methods when the subspaces are of Gaussian shape, but does substantially challenge them and is less computationally taxing than these methods. SICA can, therefore, be considered an attractive alternative to the other methods. It performs quickly and outperforms them when the true components are multimodal or, in the case of unimodal, Gaussian components, challenges them.

5.1 Limitations and future directions

Limitations to this thesis were that no method for model selection was used in the simulation studies. Instead, for every method and factor combination, the true number of components and clusters were used. This is not akin to real-life situations since these require model selection in order to obtain the final combination of number of components and clusters. As described in Durieux and Wilderjans (2019), an easy and intuitive method for deciding the number of components for SICA is choosing the number of components that exceed the 0.555 threshold for the Bimodality Coefficient (SAS, 2004). Consequently, the number of clusters can be decided using existing formal model selection pro-

cedures such as CHull (Ceulemans & Kiers, 2006; Wilderjans, Ceulemans, & Meers, 2013). CHull finds the optimal model, balancing model fit and complexity, from a range of fitted models and allows to compare models with, in the context of cluster analysis, different numbers of components and/or clusters.

A second limitation is that all compared methods use some form of *linear* transformation (i.e., dimension reduction) to map the full data on to a lower-dimensional subspace. Methods that apply *non-linear* transformations were not taken into account, but may very well be of interest to the reader. The exercise of finding non-linear representations of data to which cluster analysis can be applied is called manifold clustering, and we refer the interested reader to Abdolali and Gillis (2021) for a brief overview of manifold clustering methods. In addition, this thesis has also not considered the exercise of clustering *sparse* data and we refer to Witten and Tibshirani (2010) for an example of a method for doing so.

A third limitation, also described in Steinley et al. (2012) and Durieux and Wilderjans (2019), is that the sequential and/or orthogonal fashion in which the described methods extract components resulting in single, unidimensional representations of cluster structures. It is, however, entirely conceivable that cluster structures are not well representable in such unidimensional component extractions and may, instead, require multidimensional representations to be held intact. Such representations have not been discussed in this thesis and new research is required to extrapolate the univariate nature of the BC criterion to a larger number of dimensions.

Lastly, all methods applied in this thesis utilise K-means clustering resulting in hard, non-overlapping clusters. However, due to the fact that SICA applies dimension reduction and clustering in two separate steps, soft or fuzzy clusterings, in which subjects can belong to more than one cluster at once, can also be

applied. Incorporation of such methods in SICA requires further study.

References

Abdolali, M., & Gillis, N. (2021). Beyond linear subspace clustering: A comparative study of nonlinear manifold clustering algorithms. *arXiv preprint arXiv:2103.10656*.

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., ... Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 270–279.

Arabie, P., & Hubert, L. (1996). Advances in cluster analysis relevant to marketing research. In *From data to knowledge* (pp. 3–19). Springer.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256.

Aron, E. N., & Aron, A. (1997). Sensory-processing sensitivity and its relation to introversion and emotionality. *Journal of personality and social psychology*, *73*(2), 345.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior research methods*, *37*(3), 379–384.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International conference on database theory* (pp. 217–235).

Bugrien, J. B., & Kent, J. T. (2009). Independent component analysis: An approach to clustering. *signs*, *98*, 8a.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Ceulemans, E., & Kiers, H. A. (2006). Selecting among three-mode principal

component models of different types and complexities: A numerical convex hull based method. *British journal of mathematical and statistical psychology*, *59*(1), 133–150.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, *36*(3), 287–314.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., ... Colliot, O. (2011). Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *neuroimage*, *56*(2), 766–781.

D'Enza, A. I., Markos, A., van de Velden, M., & Markos, M. A. (2016). Package 'clustrd'.

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

de Vos, F., Schouten, T. M., Hafkemeijer, A., Dopper, E. G., van Swieten, J. C., de Rooij, M., ... Rombouts, S. A. (2016). Combining multiple anatomical mri measures improves alzheimer's disease classification. *Human brain mapping*, *37*(5), 1920–1929.

de Vos, F., Schouten, T. M., Koini, M., Bouts, M. J., Feis, R. A., Lechner, A., ... Rombouts, S. A. (2020). Pre-trained mri-based alzheimer's disease classification models to classify memory clinic patients. *NeuroImage: Clinical*, *27*, 102303.

Durieux, J., & Wilderjans, T. F. (2019, February). *Subspace partitional clustering obtained by independent component analysis.*

Feng, C., Liu, S., Zhang, H., Guan, R., Li, D., Zhou, F., ... Feng, X. (2020). Dimension reduction and clustering models for single-cell rna sequencing data: A comparative study. *International journal of molecular sciences*,

21(6), 2181.

Forgey, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, *21*(3), 768–769.

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior research methods*, *45*(1), 83–97.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, *100*(9), 881–890.

Gultepe, E., & Makrehchi, M. (2018). Improving clustering performance using independent component analysis and unsupervised feature learning. *Human-centric Computing and Information Sciences*, *8*(1), 1–19.

Handels, R. L., Aalten, P., Wolfs, C. A., OldeRikkert, M., Scheltens, P., Visser, P. J., ... Verhey, F. R. (2012). Diagnostic and economic evaluation of new biomarkers for alzheimer's disease: the research protocol of a prospective cohort study. *BMC neurology*, *12*(1), 1–8.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *Annals of statistics*, *13*(1), 70–84.

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

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193–218.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, *10*(3), 626–634.

Hyvärinen, A., & Kano, Y. (2003). Independent component analysis for non-normal factor analysis. In *New developments in psychometrics* (pp. 649–

656). Springer.

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis.* John Wiley and Sons.

Jansen, W. J., Handels, R. L., Visser, P. J., Aalten, P., Bouwman, F., Claassen, J., ... Ramakers, I. H. (2017). The diagnostic and prognostic value of neuropsychological assessment in memory clinic patients. *Journal of Alzheimer's Disease*, *55*(2), 679–689.

Lawrence, M. A. (2016). Package 'ez'. *R package version*, *4*(0).

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., ... Phelps, C. H. (2011). The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 263–269.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *psychometrika*, *45*(3), 325–342.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. *Psychometrika*, *50*(1), 123–127.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179.

Nascimento, M., Silva, F. F. e., Sáfadi, T., Nascimento, A. C. C., Ferreira,

T. E. M., Barroso, L. M. A., ... Serão, N. V. L. (2017). Independent component analysis (ica) based-clustering of temporal rna-seq data. *PloS one*, *12*(7), e0181195.

Parsons, L., Haque, E., & Liu, H. (2004). Subspace clustering for high dimensional data: a review. *Acm sigkdd explorations newsletter*, *6*(1), 90–105.

Qiu, W., & Joe, H. (2006a). Generation of random clusters with specified degree of separation. *Journal of Classification*, *23*(2), 315–334.

Qiu, W., & Joe, H. (2006b). Separation index and partial membership for clustering. *Computational statistics & data analysis*, *50*(3), 585–603.

Qiu, W., Joe, H., & Qiu, M. W. (2006). The clustergeneration package.

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

SAS, R. (2004). *Sas/stat users guide. release 9.1.* SAS Institute Cary, NC.

Seiler, S., Schmidt, H., Lechner, A., Benke, T., Sanin, G., Ransmayr, G., ... Schmidt, R. (2012). Driving cessation and dementia: Results of the prospective registry on dementia in. *PLoS ONE*, *7*(12).

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Smolewska, K. A., McCabe, S. B., & Woody, E. Z. (2006). A psychometric evaluation of the highly sensitive person scale: The components of sensory-processing sensitivity and their relation to the bis/bas and "big five". *Personality and Individual Differences*, *40*(6), 1269–1279.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In *New directions in statistical physics* (pp. 273–309). Springer.

Steinley, D. (2003). Local optima in k-means clustering: what you don't know

may hurt you. *Psychological methods*, *8*(3), 294.

Steinley, D., & Brusco, M. J. (2008). A new variable weighting and selection procedure for k-means cluster analysis. *Multivariate Behavioral Research*, *43*(1), 77–108.

Steinley, D., Brusco, M. J., & Henson, R. (2012). Principal cluster axes: A projection pursuit index for the preservation of cluster structures in the presence of data reduction. *Multivariate behavioral research*, *47*(3), 463–492.

Timmerman, M. E., Ceulemans, E., De Roover, K., & Van Leeuwen, K. (2013). Subspace k-means clustering. *Behavior research methods*, *45*(4), 1011–1023.

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

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

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

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). Chull: A generic convex-hull-based model selection method. *Behavior research methods*, *45*(1), 1–15.

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, *105*(490), 713–726.

Yeung, K. Y., & Ruzzo, W. L. (2001). Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on

principal component analysis for clustering gene expression data. *Bioinformatics*, *17*(9), 763–774.

Appendices

A Code

All code and data necessary to reproduce this thesis will be made available on Git and accessible to the readers via $https://github.com/LJRozema/MSmasterthesis.git$.

B Interaction effects multimodality simulation 1

B.1 *BC* values

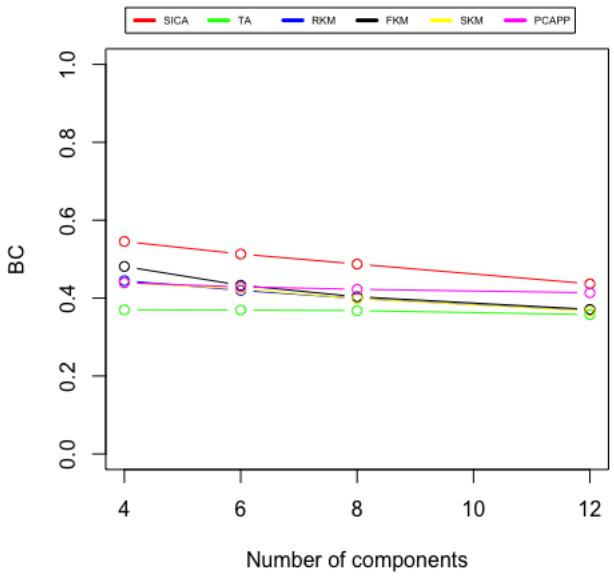

Figure B.1. Average median *BC* values per method and number of components making up the cluster space. *BC* values deteriorate, except for TA, for which it remains the same.

There were interaction effects between method and the number of components ($\eta^2_G = 0.34$, $F(5, 600) = 104.67$, $p < .001$) and method and number of clusters ($\eta^2_G = 0.19$; $F(5, 600) = 48.17$, $p < .001$). The first interaction effect is visualised in Figure B.1. From Figure B.1, it becomes apparent that, for all methods, *BC* values decrease as the number of components increase and that

this trend trend is linear for all methods but especially pronounced for SICA. PCAPP and TA, on the other hand, appear to be invariant to changes in BC values due to the number of components making up the subspace.

The interaction effect between method and number of clusters ($\eta^2_G = 0.19$; $F(5,\ 600) = 48.17$, $p < .001$) present in the data is visualised in Figure B.2. Inspecting Figure B.2, it becomes apparent that BC values slightly increase as the number of cluster present in the data increases. This could be explained by the fact that more clusters in the data requires more components that show multimodality, making for increasing BC values in the methods aiming to extract multimodal components.

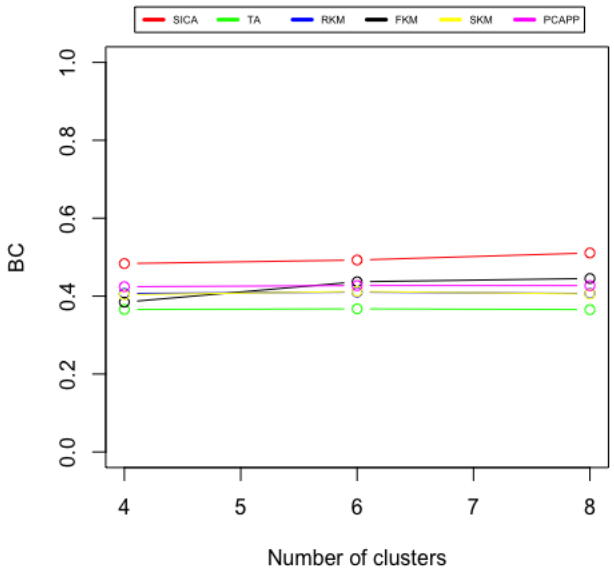

Figure B.2. Average median BC values per method and number of clusters present in the data. BC values slightly increase as the number of clusters increase.

B.2 Dip test *p*-values

There was a substantial interaction effect between method and number of components defining the cluster subspace ($\eta^2_G = 0.30$, $F(5, 600) = 85.95$, $p < .001$). Inspecting Table 4 and Figure B.3, it becomes clear that all methods show increasing average median p-values as the number of components defining the subspace increases. SICA, despite showing the most bimodal components of all methods, demonstrates this trend too, but does so steeper than the other methods. This can be explained by the fact that the other methods appear, almost by default, to extract Gaussian shapes and, in doing so, are invariant to the number of components.

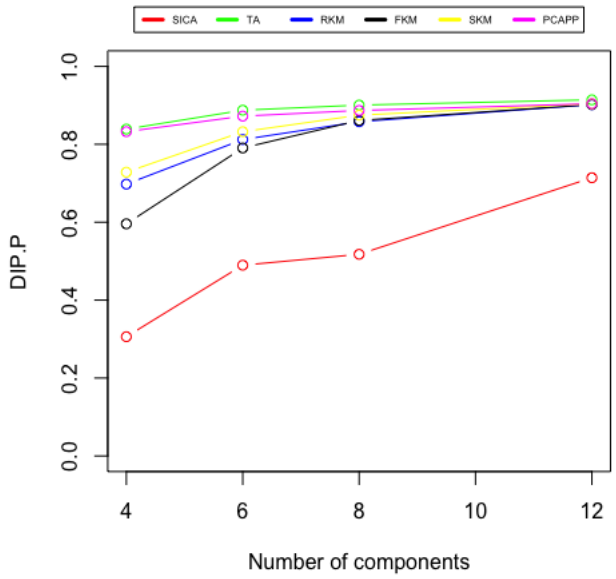

Figure B.3. Average median Dip test p-values per method and number of components. Dip test p-values increase as the number of components defining the subspace increase.

A second interaction effect was found between method and the number of masking variables present in the data ($\eta^2_G = 0.16$; $F(5, 600) = 39.56$, $p < .001$). Inspecting Figure B.4, it is apparent that most methods remain invariant to increasing numbers of noise variables when it comes to their average median p-values. SICA, conversely, demonstrates a trend that it extracts components from the data that show higher multimodality as the number of noise variables increases. There is no clear explanation for this observation and it is contrary to expectations.

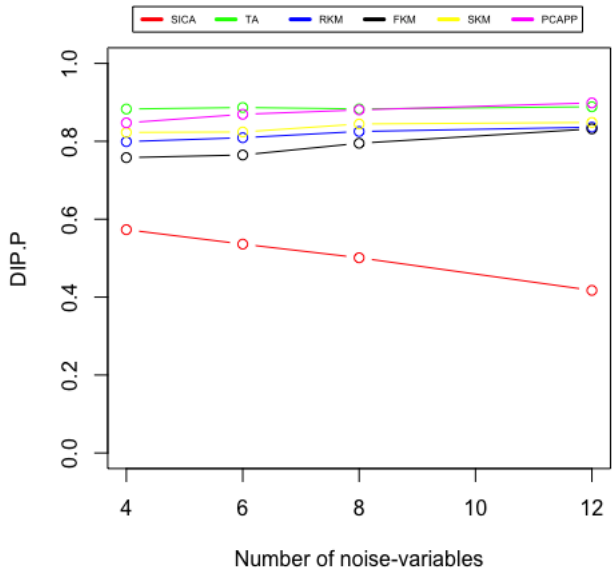

Figure B.4. Average median Dip test p-values per method and number of noise variables. Dip test p-values increase as the number of components defining the subspace increase, except for SICA, for which they decrease.

The last interaction was found between method and the division of subjects over clusters ($\eta^2_G = 0.28$; $F(10, 120) = 39.92$, $p < .001$) and is shown in Figure

B.5. In this figure, it shows that all methods except SICA are not very sensitive to the way in which subjects are divided over clusters for their average median p-values. SICA, again, shows a different trend and does appear sensitive to this division because it extracts components demonstrating less multimodality for the large condition.

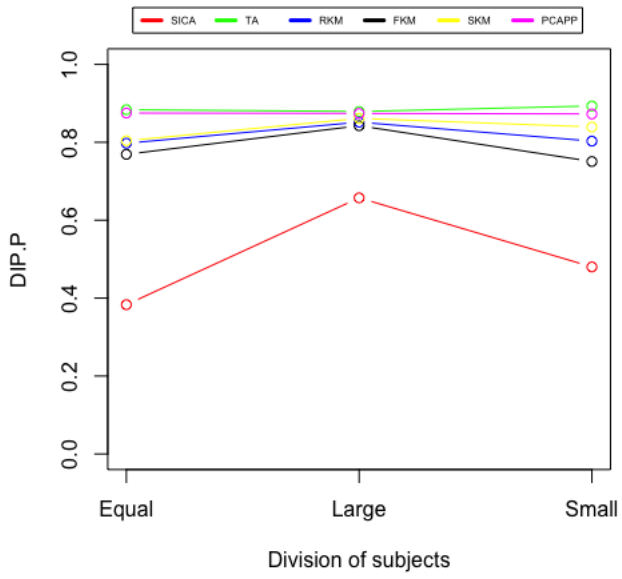

Figure B.5. Average median Dip test p-values per method and division of subjects over clusters. SICA shows the lowest average median Dip test p-values of all methods but these increase for the large level.

C Interaction effects multimodality simulation 2

C.1 *BC* values

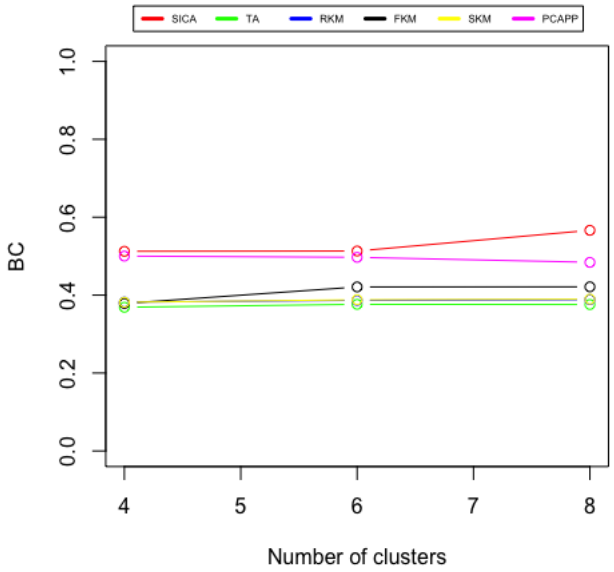

Figure C.1. Average median *BC* values per method and number of clusters present in the data. SICA shows increased multimodality as the number of clusters increases.

Interaction effects were found between method and each of the four factors, each of which will be discussed. First, the number of clusters present in the data substantially differed per method ($\eta^2_G = 0.20$, $F(5, 600) = 39.38$, $p < .001$). Inspecting Table 8 and Figure C.1 reveals that FKM and SICA demonstrate a small increase in average median *BC* values of the extracted components when

the number of clusters increases, whereas other methods show little effect. Interestingly, this simulation shows that PCAPP extracts somewhat less multimodal components when larger numbers of clusters are present, despite the fact that it maximises a metric quantifying multimodality.

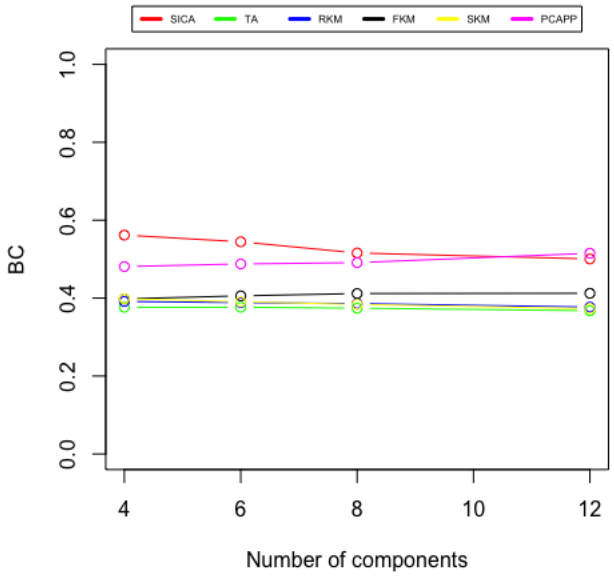

Figure C.2. Average median BC values per method and number of components. SICA shows decreased multimodality as the number of components increases.

The influence of the number of components defining the cluster subspace also differed per method ($\eta^2_G = 0.25$; $F(5, 600) = 51.95$, $p < .001$), as shown in Figure C.2. Again, all PCA-based methods show little to no effect of the number of components on the multimodality of their extracted components, shown by the flat lines for these methods. SICA shows a somewhat decreasing trend, meaning BC values decrease as the number of components defining the cluster subspace increase. In contrast, PCAPP shows a minor increase in its

ability to identify multimodal components for larger numbers of components, depicted by the increasing pink line.

The influence of the number of noise variables added to the data differed per method ($\eta^2_G = 0.28,\ F(5,\ 600) = 60.91,\ p < .001$). As shown in Figure C.3, FKM and PCAPP are somewhat less capable of identifying non-Gaussian components as the number of noise variables increases as their lines decrease. SICA demonstrates invariance regarding multimodality as the number of noise variables increases, whereas other methods show diminished multimodality

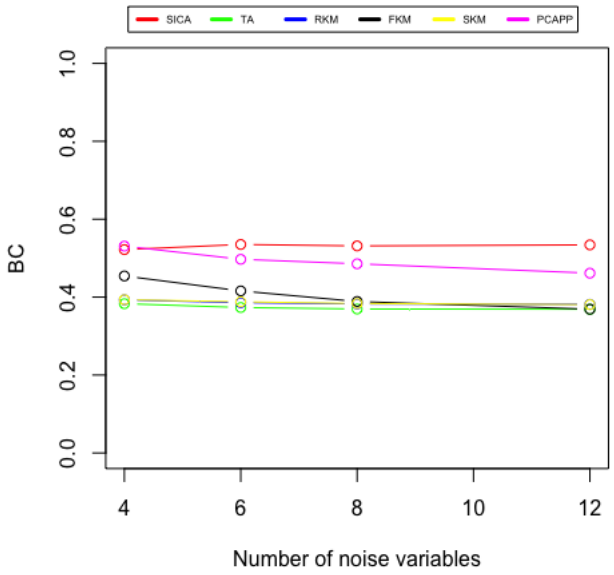

Figure C.3. Average median *BC* values per method and number of noise variables.

Lastly, the effect of the division of subjects over clusters on the capacity of the models to extract multimodal components differed substantially per method ($\eta^2_G = 0.20,\ F(10,\ 600) = 19.13,\ p < .001$). Figure C.4 reveals that this is

only really the case for SICA. SICA shows a minor decrease in average median BC value of its extracted components for the case in which the cluster space is dominated by a large cluster, compared to when equal clusters or a combination of one small cluster and equal remaining clusters make up the cluster structure. All other methods show to be invariant to the division of subjects over clusters and its effect of the multimodality of its extracted components.

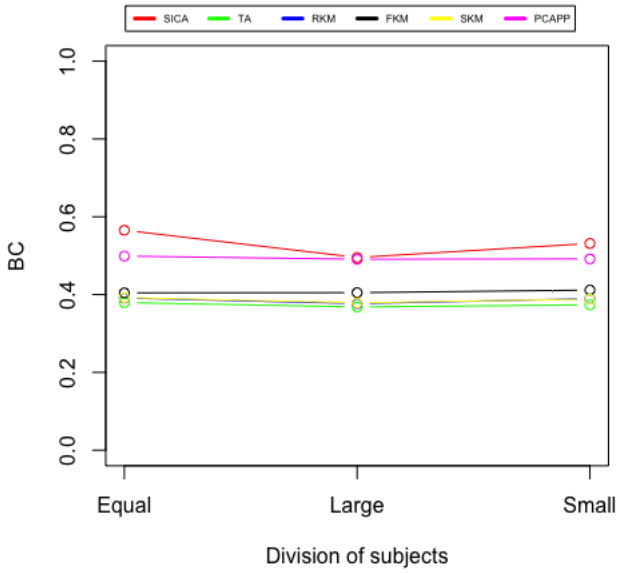

Figure C.4. Average median BC values per method and division of subjects over clusters.

C.2 Dip test *p*-values

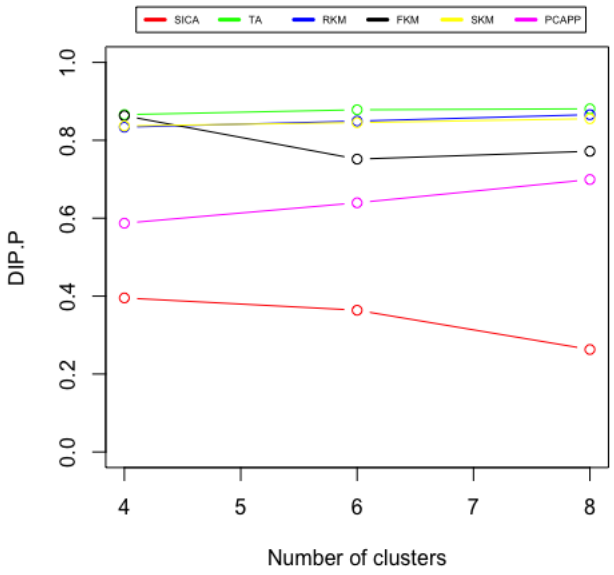

Figure C.5. Average median Dip test p-values values per method and number of clusters present in the data. SICA components show increased multimodality as the number of clusters increases.

The interaction between method and number of clusters in the data showed a η^2_G value of $\eta^2_G = 0.22$ ($F(5, 600) = 41.45$, $p < .001$) and is presented visually in Figure C.5. All methods except FKM and SICA show an overall increase in average median p-values as the number of clusters present in the data increases. SICA and FKM show a decrease in average median p-values as this number increases. For SICA, this is conform expectations as larger numbers of clusters should make for more multimodal components and SICA identifies this pattern. For FKM, this pattern in surprising as it is based on PCA, which should be

relatively invariant to the non-Gaussianity of the components making up the subspace. It should be noted that, possibly due to taking the average median, that average median p-values in Table 9 do not, for any method, fall below the standard $\alpha = 0.05$ to conclude formal significance of the Dip test.

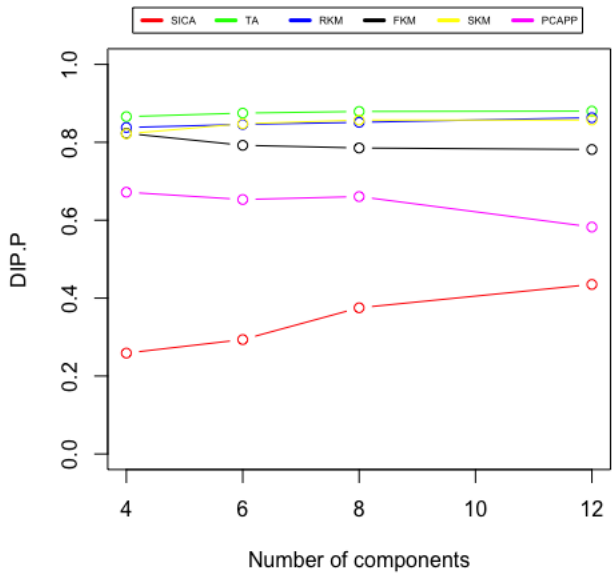

Figure C.6. Average median Dip test p-values values per method and number of components. SICA components show decreased multimodality as the number of components increases.

The interaction between method and number of components making up the subspace also demonstrated to be substantial $\eta^2_G = 0.19$ ($F(5, 600) = 35.36$, $p <$ .001). Figure C.6 shows this interaction visually. From the figure, it is observed that most methods are relatively invariant to change in average median p-values indicated by the flat lines as the number of components increases. SICA shows an increase in average median p-values as the number of components increases,

possibly due to the fact that it maximises non-Gaussianity for the first few extracted components, while remaining extracted components may demonstrate more Gaussian shapes. PCAPP, which also aims at identifying multimodal components, shows an opposite trend in which, as the number of components defining the subspace increases, its average median p-values *decrease*.

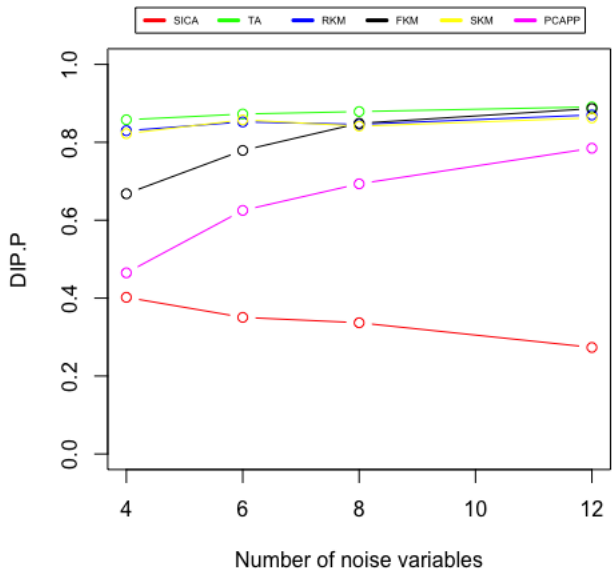

Figure C.7. Average median Dip test p-values values per method and number of noise variables present in the data. Contrary to expectations, SICA components show increased multimodality as the number of noise variables increases.

The interaction between method and number of masking variables present in the data is presented in Figure C.7 and shows the expected pattern for all methods except SICA ($\eta^2_G = 0.39$; $F(5, 600) = 96.19$, $p < .001$). Specifically, as the number of noise variables increases, average median p-values of all methods increase, meaning that it becomes harder for each of these methods to identify

non-Gaussian subspaces. This was also found in simulation 1. SICA demonstrates the unexpected trend that it manages to identify components displaying bimodality to a larger extent when the number of noise variables in the data increases.

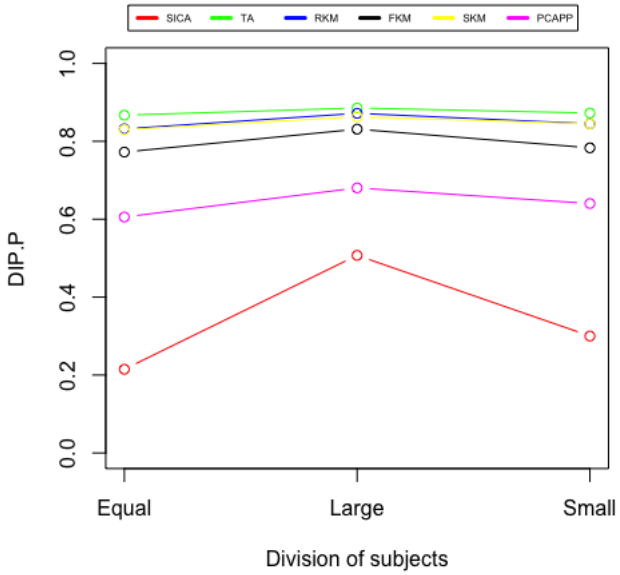

Figure C.8. Average median Dip test p-values values per method and division of subjects over the clusters. For the large level, all methods show worst multimodality.

The interaction between method and division of subjects over the clusters, lastly, had a η^2_G value of $\eta^2_G = 0.28$ ($F(10,\ 600) = 29.46$, $p < .001$) and is presented in Figure C.8. All methods show the lowest average median p-values for the equal and small divisions of subjects whereas these increase for the large division of subjects. SICA appears very sensitive to this manipulation as its average median p-values increase dramatically for the large division of subjects.

This pattern was also observed in the first simulation study.

D Model selection empirical data sets

D.1 Graz data

Table D1

ARI and Silhouette values for each method for different numbers of components from applying each method to the Graz data.

N comps	ARI SICA	Sil SICA	ARI TA	Sil TA	ARI RKM	Sil RKM	ARI FKM	Sil FKM	ARI SKM	Sil SKM	ARI PCAPP	Sil PCAPP
2	0.03	0.36	0.39	0.43	0.41	0.54	0.00	0.44	0.04	0.19	0.10	0.40
3	0.03	0.24	0.39	0.36	0.41	0.51	0.00	0.33	0.04	0.16	0.08	0.32
4	0.03	0.18	0.39	0.31	0.41	0.48	0.00	0.25	0.04	0.15	0.08	0.32
5	0.03	0.15	0.41	0.29	0.41	0.46	0.00	0.20	0.04	0.13	0.12	0.27
6	0.04	0.13	0.41	0.27	0.41	0.44	0.00	0.19	0.04	0.13	0.10	0.25
7	0.02	0.11	0.41	0.25	0.41	0.41	0.00	0.17	0.04	0.12	0.10	0.23
8	0.02	0.10	0.41	0.24	0.41	0.39	0.00	0.15	0.04	0.12	0.14	0.22
9	0.02	0.09	0.41	0.23	0.41	0.37	0.00	0.15	0.04	0.11	0.16	0.19
10	0.02	0.08	0.41	0.22	0.41	0.35	0.00	0.12	0.04	0.11	0.23	0.19
11	0.00	0.07	0.41	0.22	0.41	0.34	0.00	0.12	0.04	0.10	0.22	0.18
12	0.02	0.06	0.41	0.21	0.41	0.33	0.00	0.11	0.04	0.10	0.26	0.20
13	0.00	0.06	0.41	0.21	0.41	0.32	0.00	0.10	0.04	0.10	0.28	0.19
14	0.00	0.05	0.41	0.20	0.41	0.31	0.00	0.09	0.04	0.09	0.24	0.18
15	0.00	0.05	0.41	0.20	0.41	0.30	0.00	0.09	0.04	0.09	0.25	0.18
16	0.00	0.05	0.41	0.20	0.41	0.29	0.00	0.09	0.04	0.09	0.25	0.17
17	0.01	0.04	0.41	0.19	0.41	0.29	0.00	0.07	0.04	0.08	0.25	0.16
18	0.00	0.04	0.41	0.19	0.41	0.28	0.00	0.08	0.04	0.08	0.26	0.16
19	0.00	0.04	0.41	0.19	0.41	0.27	0.00	0.08	0.04	0.08	0.26	0.15
20	0.00	0.04	0.41	0.18	0.41	0.27	0.00	0.08	0.04	0.08	0.22	0.15

Note: Goodness of recovery and cluster validity do not substantially increase for larger number of components for any of the methods (the only exception being the ARI values for PCAPP) and so the number of components for this analysis was set to two.

D.2 LeARN data

Table D2

ARI and Silhouette values for each method for different numbers of components from applying each method to the LeARN data.

N comps	ARI SICA	Sil SICA	ARI TA	Sil TA	ARI RKM	Sil RKM	ARI FKM	Sil FKM	ARI SKM	Sil SKM	ARI PCAPP	Sil PCAPP
2	0.00	0.34	0.06	0.36	0.08	0.42	0.00	0.55	0.06	0.11	0.01	0.44
3	0.03	0.31	0.06	0.28	0.08	0.40	0.00	0.41	0.06	0.10	0.02	0.36
4	0.03	0.21	0.07	0.23	0.08	0.36	0.00	0.32	0.06	0.09	-0.00	0.25
5	0.02	0.17	0.07	0.20	0.08	0.34	0.00	0.26	0.06	0.08	-0.00	0.20
6	0.01	0.14	0.07	0.18	0.08	0.31	-0.01	0.22	0.06	0.07	-0.01	0.18
7	0.01	0.12	0.07	0.17	0.08	0.30	0.00	0.19	0.06	0.07	0.01	0.18
8	0.00	0.11	0.07	0.16	0.08	0.28	0.00	0.18	0.06	0.06	0.02	0.17
9	0.00	0.09	0.08	0.15	0.08	0.26	0.00	0.16	0.06	0.06	0.01	0.15
10	0.01	0.08	0.08	0.15	0.08	0.25	0.00	0.14	0.06	0.05	0.05	0.14
11	0.01	0.07	0.08	0.14	0.08	0.24	0.01	0.13	0.06	0.05	0.03	0.14
12	0.00	0.07	0.08	0.14	0.08	0.23	-0.01	0.12	0.06	0.04	0.03	0.13
13	0.00	0.06	0.08	0.13	0.08	0.22	-0.01	0.11	0.06	0.04	0.03	0.12
14	0.00	0.06	0.08	0.13	0.08	0.21	0.00	0.11	0.06	0.04	0.03	0.11
15	0.00	0.05	0.07	0.12	0.08	0.21	0.01	0.11	0.06	0.04	0.03	0.11
16	0.01	0.05	0.07	0.12	0.08	0.20	0.00	0.11	0.06	0.03	0.02	0.11
17	0.00	0.04	0.07	0.12	0.08	0.19	0.00	0.09	0.06	0.03	0.03	0.11
18	0.01	0.04	0.07	0.12	0.08	0.19	0.00	0.09	0.06	0.03	0.03	0.10
19	0.00	0.04	0.07	0.11	0.08	0.18	-0.01	0.08	0.06	0.03	0.02	0.10
20	-0.01	0.03	0.08	0.11	0.08	0.17	0.00	0.08	0.06	0.03	0.03	0.09

Note: Goodness of recovery and cluster validity do not substantially increase for larger number of components for any of the methods and so the number of components for this analysis was set to two.

D.3 SPS data

Table D3

Calinski-Harabasz values for each method for each number of clusters for the SPS data.

Number of clusters	SICA	TA	RKM	FKM	SKM	PCAPP
2	346.22	1177.72	1547.83	1.26	174.27	1067.45
3	364.19	985.01	1354.55	2.01	87.55	936.47
4	379.05	833.33	1179.61	2.73	60.90	956.07
5	377.91	758.04	983.50	4.06	51.97	1085.08
6	385.00	698.34	901.61	3.58	44.47	791.99

Note: For each model, the number of clusters yielding the highest Calinski-Harabasz value is underlined.

E Graz cluster figures

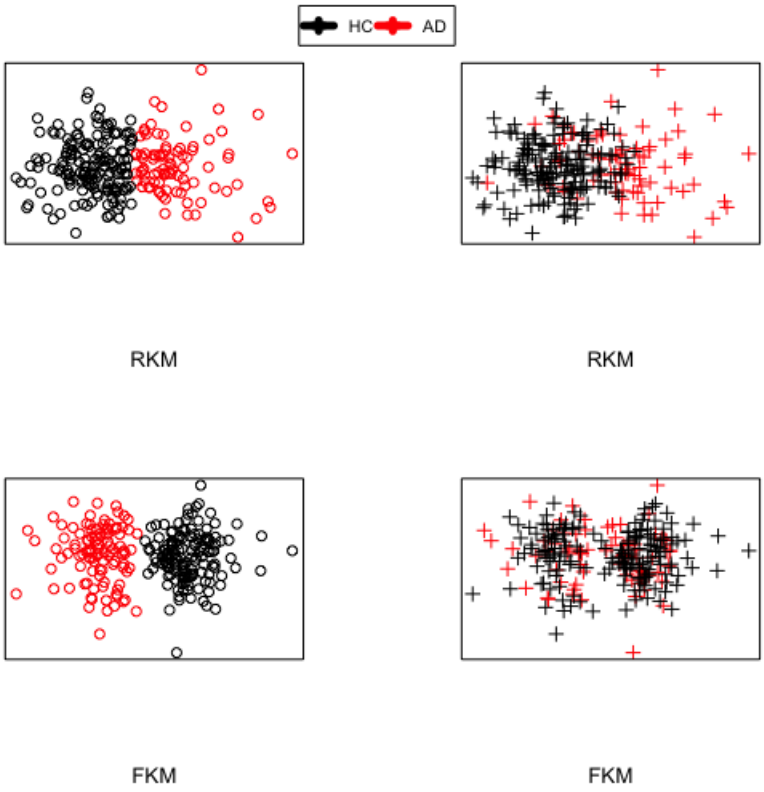

Figure E.1. Results of applying RKM and FKM to the two-cluster Graz data. Obtained clusterings are shown on the left and true diagnoses shown on the right, with the axes being the components extracted by each method.

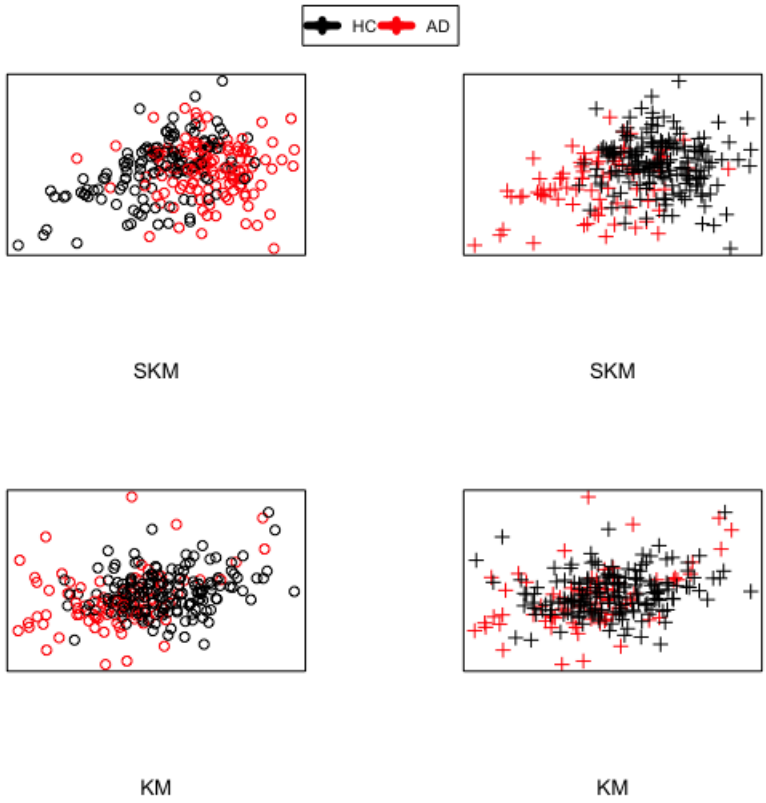

Figure E.2. Results of applying SKM and KM to the two-cluster Graz data. Obtained clusterings are shown on the left and true diagnoses shown on the right, with the axes being the components extracted by each method. Note that no dimension reduction was performed for K-means, and so the axes are the first two variables from the data set.

F LeARN cluster figures

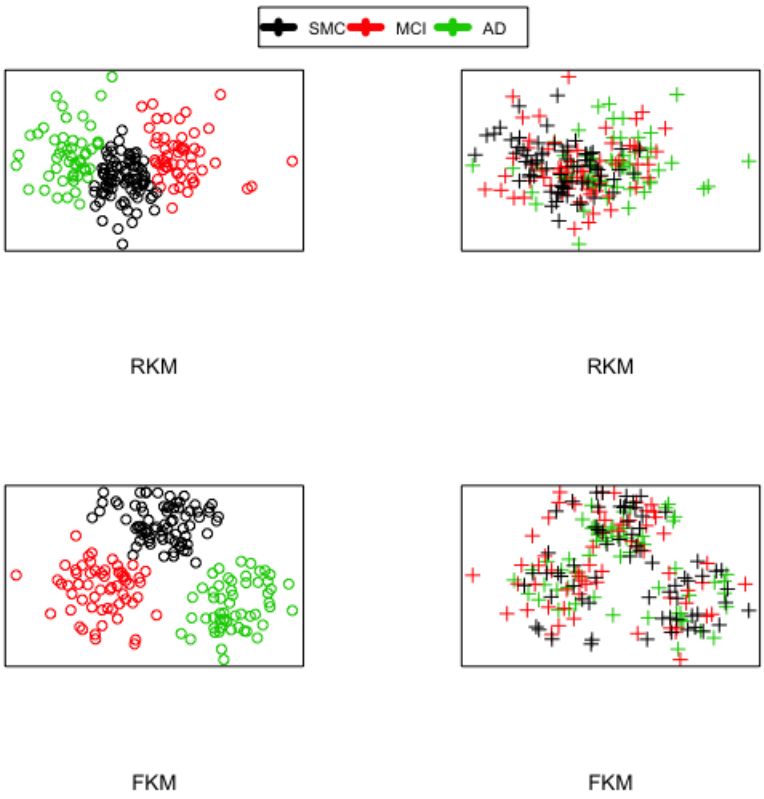

Figure F.1. Results of applying RKM and FKM to the three-cluster LeARN data. Obtained clusterings are shown on the left and true diagnoses shown on the right, with the axes being the components extracted by each method.

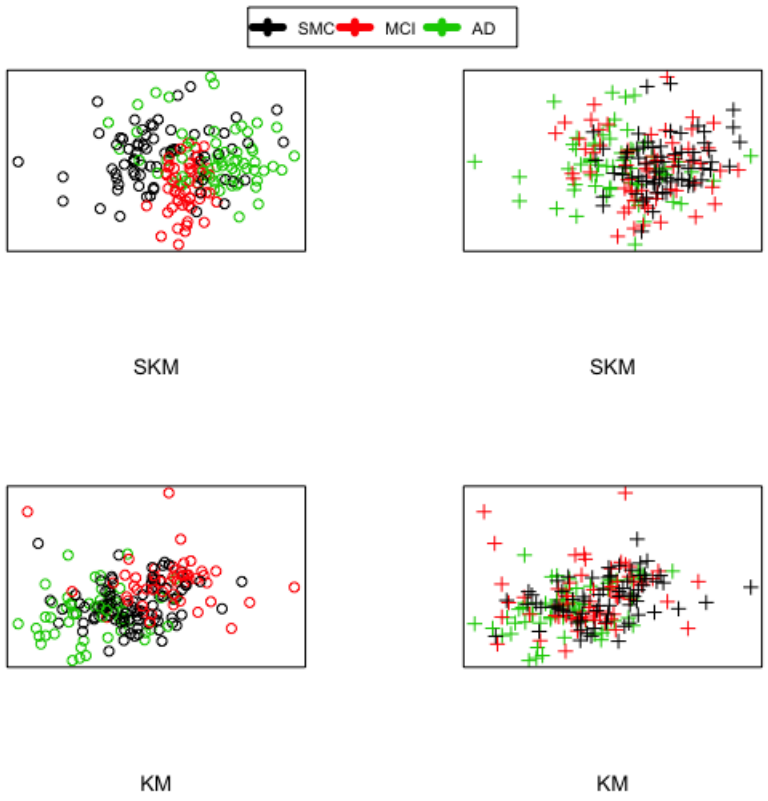

Figure F.2. Results of applying SKM and KM to the three-cluster LeARN data. Obtained clusterings are shown on the left and true diagnoses shown on the right, with the axes being the components extracted by each method. Note that no dimension reduction was performed for K-means, and so the axes are the first two variables from the data set.

G Silhouette plots SPS data

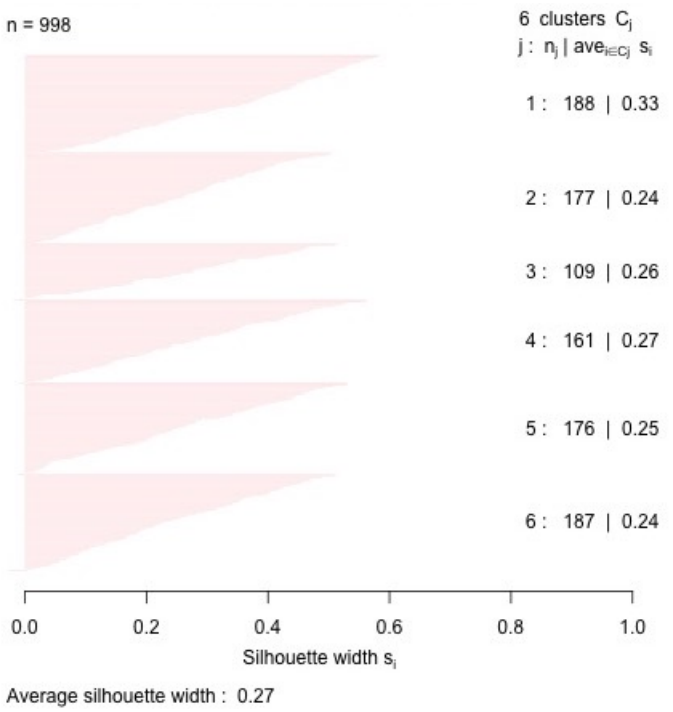

Figure G.1. Silhouette plots of the application of SICA to the SPS data with six clusters.

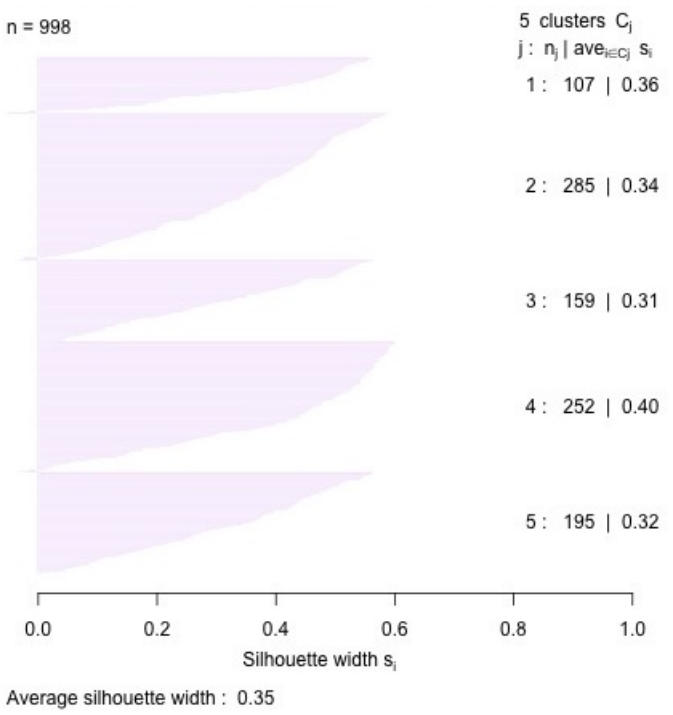

Figure G.2. Silhouette plots of the application of PCAPP to the SPS data with five clusters.

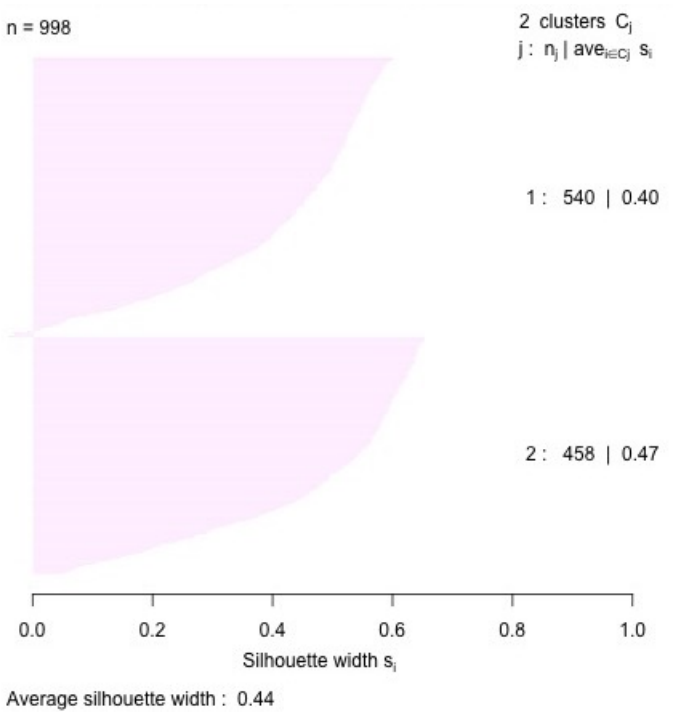

Figure G.3. Silhouette plots of the application of TA to the SPS data with two clusters.

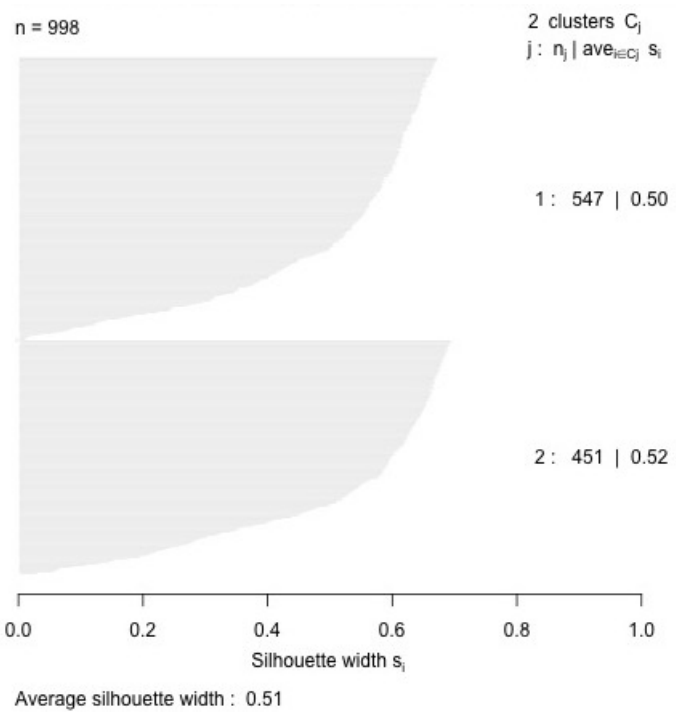

Figure G.4. Silhouette plots of the application of RKM to the SPS data with two clusters.

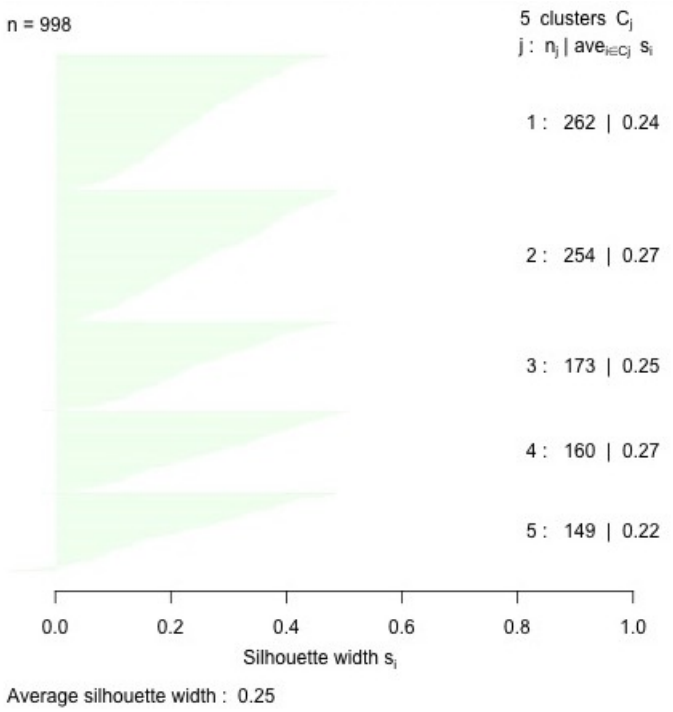

Figure G.5. Silhouette plots of the application of FKM to the SPS data with five clusters.

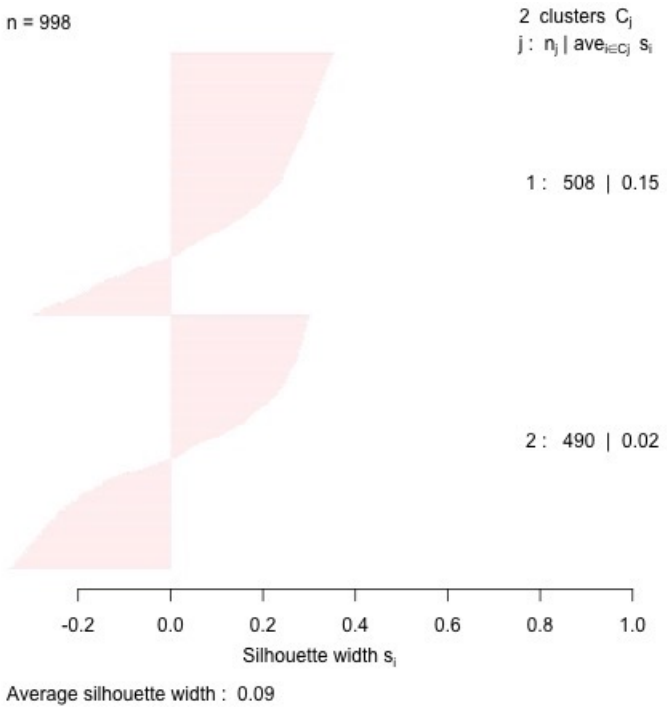

Figure G.6. Silhouette plots of the application of SKM to the SPS data with two clusters.

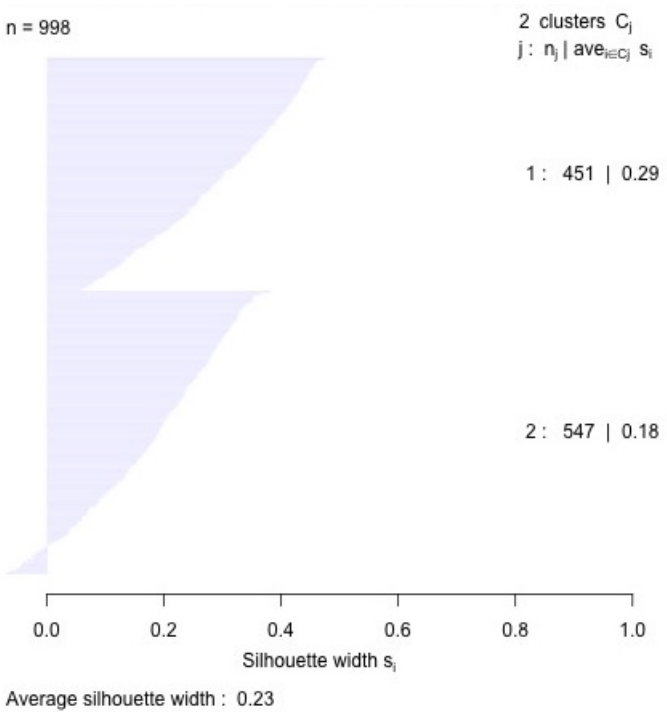

Figure G.7. Silhouette plots of the application of K-means to the SPS data with two clusters.