Investigating model order selection methods for Clusterwise ICA (C-ICA)

Boes, Amande

Universiteit Leiden

Faculteit der Sociale Wetenschappen

Investigating model order selection methods for Clusterwise ICA (C-ICA)

Amande Boes

Master’s Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioural Sciences, Leiden University

Date: February 22, 2022

Supervisor: Dr. T.F. Wilderjans and J. Durieux MSc

Abstract

Clusterwise independent component analysis (C-ICA) is a new and promising unsupervised learning method which clusters patients based on differences in spatial functional connectivity (sFC) patterns between these a priori unknown clusters of patients. When performing C-ICA, the number of clusters and components needs to be specified. However, in many cases there is no a priori information about the optimal number of clusters and components to select for. Thereby, various (mainly sequential) model order selection methods were proposed during the last years. This thesis investigated several simultaneous and sequential methods to tackle the model order selection problem for C-ICA and compared them using an extensive simulation study. Overall, CHull outperformed other simultaneous model order selection methods. Only when the number of underlying components was equal to 25, a combination of CHull with AIC, AICc, BIC, KIC or MDL performed slightly better, but this increase in performance was minimal. Nevertheless, a sequential method using scree-ratios and the variance accounted for (VAF), outperformed all simultaneous methods. In this thesis, the most efficient order in this sequential method was first selecting the number of clusters and subsequently the number of components. Although it is expected that this method will work well in other studies as well, the optimal order may differ depending on the situation. Regarding the effect of the underlying number of components and clusters a dataset consists of, there was no univocal pattern for the different methods. Conversely, the effect of noise was similar for all methods: the lower the amount of noise, the better the estimation error (and vice versa).

Keywords: C-ICA, model order selection, resting-state fMRI, clusters, components

Table of content

I. Introduction

Unsupervised learning techniques such as clustering and dimension reduction are commonly used nowadays (Saxena et al., 2017). Grouping (or clustering) objects (e.g., subjects) is used in multiple disciplines, for example in science, humanities and medical science (Saxena et al., 2017). In psychology, patients with mental diseases are often allocated to a cluster (representing a diagnostic label) based on information regarding the severity and type of symptoms which are related to these mental diseases (Luteijn & Barelds, 2013). Frequently, those allocations are based on questionnaire data and neuropsychological tests instead of on neural and/or biological information which can be assumed to be more closely related to the etiology of a disease. Indeed, most of the time only indirect measurements are used that are 'symptoms' of the disease but that are not directly related to the underlying source of the disease (Hastie et al., 2017). Several studies tried to search for more direct measures of this underlying source which affects diseases by, for example, analyzing brain structures and functions (Zhang et al., 2021). Some studies used for example functional Magnetic Resonance Imaging (fMRI) data for clustering people and looked for differences in (functional) connectivity networks between subject clusters (Drysdale et al., 2017).

Heterogeneity (between subjects) in spatial functional connectivity (sFC) patterns - which can be estimated from fMRI - seems to be a relevant indicator for clinical diseases such as Alzheimer (Gili et al., 2011; Zhang et al., 2021). sFC refers to the synchrony (i.e., correlation) of activity in distinct regions of the brain, which is represented by sFC networks/patterns. Here, activity is measured by measuring the blood oxygen level dependent (BOLD) signal of fMRI data (Fox & Raichle, 2007). However, these sFC networks/patterns cannot be directly observed in data. To uncover these patterns from the data, researchers often apply a dimension reduction technique like Independent Component Analysis (ICA). ICA is a dimensionality reduction technique which reduces the data to smaller sets of independent (and non-Gaussian) components. Such a component in fMRI data represents a set of (possibly distinct) brain regions that show synchronized activity (Beckmann et al., 2005). Moreover, ICA estimates a time course for each component or sFC pattern that indicates the strength of activity of the corresponding sFC pattern at each time point.

Research showed that there is a lot of heterogeneity in sFC patterns and associated time courses of activation in patients' resting state fMRI (Zhang et al., 2021). Resting state fMRI (rs-FMRI) measures the brain activity patterns of patients with the absence of tasks (Biswal et al., 1995; Fox & Raichle, 2007). Those rs-fMRI patterns are consistently found in many studies and for many subjects; an example of such a pattern is the default mode network (Barkhof et al., 2014). Variations or disruptions in these functional connectivity networks seem to be correlated with different mental diseases like, as mentioned before, Alzheimer (Gili et al., 2011), but also Major Depression Disorder (Greicius et al., 2007), Parkinson's Disease (Olde Dubbelink et al., 2013) and schizophrenia (Lynall et al., 2010). As such, clustering subjects based on sFC patterns may give additional insights into these mental diseases (e.g., finding subtypes of these known diseases). To this end, Clusterwise Independent Component Analysis (C-ICA) was proposed. C-ICA is a new unsupervised learning method which identifies those differences in sFC patterns between a priori unknown clusters of patients (Durieux et al., 2021). The method is used to automatically cluster patients in such homogeneous groups which are unknown a priori based on (differences in) sFC patterns which are derived from ICA.

An important question for every dimensionality reduction method is the number of components one should extract. For example, for Principal Component Analysis (PCA) the optimal number of principal components for a data set at hand needs to be identified (Jolliffe, 2005). For PCA there are several ways to determine the number of components, for example by looking for an "elbow" in a scree plot (James et al., 2017), which is rather subjective, or using the eigenvalue greater than 1.0 rule (Kaiser, 1960). There exist also methods which are more objective like, the CHull method (Wilderjans et al., 2013) and Parallel analysis (Franklin et al., 1995). For the probabilistic version of PCA (pPCA), which is based on a Gaussian latent variable model (Tipping & Bishop, 1999), for example, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to address the model order selection problem (Bouveyron et al., 2011). It is hard to choose a single method that is optimal in all cases as the performance of a method is dependent on the area of interest and on the data set (and its characteristics) on which the dimension reduction will be performed (James et al., 2017).

Similar to dimensionality reduction, with clustering methods, the optimal number of clusters for a data set needs to be determined. An example of such a clustering method is K-means clustering. Similar as in PCA, an "elbow method" could be used for determining the number of clusters here (Kodinariya & Makwana, 2013). In the context of K-means, more methods are proposed in the literature for selecting the optimal number of clusters such as, information criteria AIC and BIC (Kodinariya & Makwana, 2013) and the Gap statistic (Tibshirani et al., 2001). However, there is not one best method to tackle the model order selection problem for each data set.

This short overview shows that for dimensionality reduction and clustering techniques there is relatively much known about (optimal) techniques for the selection of the number of components and the number of clusters. Much less is known about techniques for model order selection methods for models like C-ICA, where both the number of components and the number of clusters needs to be selected. Of course, it is always possible to treat both selection problems sequentially, although, it is not clear which of both problems in that case should be solved first. This is not a trivial issue as both problems may influence each other. Indeed, first selecting a wrong number of components will have a detrimental effect on the subsequent selection of the number of clusters and vice versa. In this regard, the study of De Roover et al. (2011) is interesting. In this study, a technique called clusterwise simultaneous component analysis (SCA-ECP) is introduced, for which similar to C-ICA, the optimal number of underlying components (Q) and number of clusters (R) needs to be identified. This study tackled this problem by fitting the clusterwise SCA-ECP model with multiple combinations of values for R and Q and next applying a sequential model order selection technique (De Roover et al., 2011). First, the scree-ratios which are based on the variance accounted for are averaged to find the optimal R. Subsequently, conditional on the optimal R, the model for which adding more components does not result in a significant increase in fit is chosen to be the optimal Q. An alternative solution could be to use a model order selection method that solves both selection problems (i.e., number of components and clusters) simultaneously (see Section 2.3).

As mentioned before, when performing C-ICA, one always needs to determine the optimal number of clusters R and components Q (Durieux et al., 2021). When using fMRI data and spatially independent components are sought for, Q can range from one to the number of timepoints of a subject in a dataset (i.e., which easily could be 150 or more). Selecting an incorrect number of components can negatively influence the results. Indeed, when choosing a too small number of components there is a high probability that very broad components are retained that are not easy to link to relevant neural functioning. When choosing a too large number of components, it can result in overfitting and the extraction of scarce independent components that are not related to brain functioning at all (Särelä & Vigário, 2003).

When using C-ICA, the number of clusters R can range from one to the total number of subjects. When choosing the maximum number of clusters, every individual will represent one cluster, which boils down to performing ICA on each subject’s dataset separately. But when choosing only a single cluster, everyone will be assigned to the same group, which is equal to performing a group-ICA (Calhoun et al., 2009). Note that the latter analysis is not very informative for our purpose as it will totally ignore the heterogeneity between subjects in sFC patterns. Also, the former analysis is not very informative as systematic differences between sFC patterns across subjects cannot be identified in a clear way.

Summarized, selecting the optimal number of clusters and components is not a simple task, especially because most of the time there is no a priori information about what the optimal number of components or clusters should be. Thereby, many different model selection methods were proposed throughout the years, without clear guidelines about which method to use in which situation. Also for C-ICA, in which both the number of components and clusters need to be determined, no such guidelines exist.

In this thesis, therefore, the model order selection problem for C-ICA will be tackled. The main question that will be central in this thesis is: Which model selection method is the most effective in determining the number of clusters and components in a C-ICA analysis of, for example, a fMRI dataset. The goal of this thesis is to compare multiple model order selection methods and investigate which method works best for C-ICA. Besides determining which method works best, in this thesis, sequential and simultaneous model order selection methods will also be compared, to see which of the two classes of methods is most efficient. In a sequential method, the number of components and clusters will be determined one after the other, whereas in a simultaneous method, the number of components and clusters are determined both at once. Regarding the comparison of sequential and simultaneous methods, it is expected that sequential model selection methods will increase estimation errors compared to simultaneous methods, because when first determining a wrong number of components (or clusters), this mistake may negatively influence the determination of the number of clusters (or components) in the subsequent analysis.

The remaining parts of this thesis will investigate and compare multiple model selection methods for determining the optimal number of clusters and components for a C-ICA analysis of a fMRI dataset. In the Methodology Section, all different model selection methods will be explained. In the next two sections, all those methods will be compared in an extensive simulation study to investigate which method is most effective in selecting the true model order in different situations. In Section 3, the design and procedure for the simulation study will be sketched and the results will be presented in Section 4. Finally, a conclusion and discussion of the results will be provided in Section 5, including limitations and final remarks.

II. Methodology

2.1 Data

To investigate the model order selection problem for C-ICA, multiple subjects' rs-fMRI data $\boldsymbol{X_i}$ (T time courses × V voxels) will be analysed (where $i = 1, \dots, I$ subjects). For each subject's data, the rows represent timepoints (T) and the columns the voxels (V). An example dataset will be used to illustrate some of the methods discussed later in this paper. This small example dataset consists of 2 clusters, 5 subjects per cluster, 100 voxels, 50 timepoints and 4 components (and a low amount of noise of 10%). So each of the ten subjects has a data set with 100 rows and 50 columns containing the (rs-)fMRI data. This data set will be used as a small example to demonstrate the different techniques for model order selection. To this end, C-ICA was fitted with 10 multistarts to the data, varying the number of components Q from 1 to 5 and the number of clusters R from 1 to 5. Note that the true value for $Q = 4$ and for $R = 2$.

2.2 Clusterwise Independent Component Analysis (C-ICA)

The main goal of C-ICA is to cluster subjects in homogeneous groups based on similarities and differences in underlying functional networks and their associated time courses (Durieux et al., 2021). C-ICA decomposes each $\boldsymbol{X_i}$ as:

$$\boldsymbol{X_i} = \sum_{r=1}^{R} p_{ir} \boldsymbol{A_i} \boldsymbol{S^r} + \boldsymbol{E_i} \tag{1}$$

This model shows a dataset of a single subject $\boldsymbol{X_i}$, where p_{ir} is an element of the partition matrix (I x R) which equals one if person i is allocated to cluster r, otherwise zero (Durieux et al., 2021). $\boldsymbol{S^r}$ (*Q* Components × *V* voxels) indicates the functional FC patterns matrix, where each row is an independent component or sFC pattern, for cluster R. $\boldsymbol{A_i}$ (*T* time course × *Q*) represents a subject specific mixing matrix. Finally, $\boldsymbol{E_i}$ $(T \times V)$ contains an error term for each matrix.

The goal of C-ICA is to minimize the following loss function:

$$L = \sum_{i=1}^{I} ||\boldsymbol{X_i} - \sum_{r=1}^{R} p_{ir}\boldsymbol{A_i}\boldsymbol{S^r}||^2 \tag{2}$$

In order to minimize this loss function an Alternating Least Squares (ALS) algorithm is used. This algorithm consists of three steps. It starts by randomly initializing a partition matrix $\boldsymbol{P}$. In particular, each subject will be assigned to a cluster (with equal probabilities for every cluster) and this such that there are no empty clusters (Durieux et al., 2021). Using this starting partition with about equally sized clusters, in step 2, C-ICA parameters (i.e., cluster specific $\boldsymbol{S^r}$and subject specific time courses $\boldsymbol{A_i}$) are estimated. During this second step, for every cluster, a spatial (group) ICA will be performed to the concatenated data set (i.e., data of all subjects belonging to the cluster are concatenated in the time dimension). This results in cluster specific sFC patterns $\boldsymbol{S^r}$ and subject specific time courses $\boldsymbol{A_i}$ for the subjects of that cluster. The third step includes updating the partition matrix $\boldsymbol{P}$ given the updated ICA parameters per cluster (in Step 2). This update is performed subject by subject in such a way that each subject's data block is reassigned to its optimal (i.e., best fitting) cluster. Step 2 and step 3 are repeated until the convergence criterion is achieved (i.e., decrease in loss smaller than .000001). Note that this ALS algorithm is sensitive to local minima. To this end, the ALS algorithm is repeated several times with different random start using different random start partitions. Finally, the best solution across all starts is chosen as the final solution. For the example data, as described in Section 2.1, the resulting fit values for all the fitted combinations of Q and R (both going from 1 to 5) resulting from the C-ICA analyses (using 10 random starts) are displayed in Table 1.

Table 1

C-ICA loss function values (badness of fit) for the corresponding number of components (Q) and clusters (R) for each C-ICA model applied to the example data (the true model used to generate the data is indicated in bold)

Model number	Q	R	Badness of Fit-value
1	1	1	8320.00
2	1	2	6894.87
3	1	3	6798.96
4	1	4	6729.54
5	1	5	6688.78
6	2	1	6664.57
7	2	2	4571.76
8	2	3	4427.36
9	2	4	4359.67
10	2	5	4262.98
11	3	1	5476.77
12	3	2	2728.07
13	3	3	2594.21
14	3	4	2480.52
15	3	5	2447.85
16	4	1	4336.81
17	**4**	**2**	**1064.96**
18	4	3	1056.07
19	4	4	1047.90
20	4	5	1040.44
21	5	1	3341.77
22	5	2	1039.34
23	5	3	1026.96
24	5	4	1015.39
25	5	5	1004.73

2.3 Simultaneous model order selection methods

Selecting the optimal number of clusters and number of components simultaneously can be done by running a C-ICA analysis multiple times with increasing numbers of components and clusters, and at the end selecting the optimal number of R and Q by using a model selection heuristic. Most of these heuristics are based on a goodness/badness of fit value and a model complexity value for each fitted model (see, for example, Table 1).

Badness of fit. In this thesis, for every simultaneous model order selection method, one of two measures of fit are used, that is (1) either the least squares fit as calculated in Equation 2 (i.e., the loss function that is minimized by the C-ICA algorithm) or (2) the negative log-likelihood which is derived from a minimal stochastic extension of the C-ICA method (and is linked to the least squares fit). In particular, when assuming that all error terms are independent of each other and follow a normal distribution with mean zero and a variance (that has to be estimated), the negative log-likelihood is calculated as follows:

$$-l(X|M) = \frac{n}{2}\log(2\pi) + n\log(\sigma_e) + \frac{1}{2\sigma_e^2}\sum_{i=1}^{I}||\boldsymbol{X_i} - \sum_{r=1}^{R}p_{ir}\boldsymbol{A_i}\boldsymbol{S^r}||^2 \quad (3)$$

Here, $\sum_{i=1}^{I}||\boldsymbol{X_i} - \sum_{r=1}^{R}p_{ir}\boldsymbol{A_i}\boldsymbol{S^r}||^2$ is equal to the badness of fit which results from the C-ICA analysis as given in Equation (2). The error variance (σ_e) is calculated by dividing this badness of fit by the total number of elements n (which equals the sum of the number of data points - $T \times V$ - across all subjects ($i = 1, \ldots, I$).

$$\sigma_e^2 = \frac{\sum_{i=1}^{I}||\boldsymbol{X_i} - \sum_{r=1}^{R}p_{ir}\boldsymbol{A_i}\boldsymbol{S^r}||^2}{n} \quad (4)$$

Model complexity. Considering not only the goodness/badness-of-fit, but also the right model complexity is important for model selection (Myung, 2000). Different measures of complexity may lead to different results. For the C-ICA model it is yet unclear what would be an optimal measure to quantify its complexity. Taking the number of fitted parameters as the complexity for C-ICA is not correct as some parameters are dependent on each other (e.g., when for object i $p_{ir} = 1$ for cluster r, it is 0 for all other clusters). The influence of this complexity measure will be investigated by using five different complexity measures as displayed in Table 2. The first two complexity measures only look at Q and R. Complexity measures 3, 4 and 5 are extended versions of Complexity measures 1 and 2 and drafted in such a way that they also take somehow into account the size of the data (i.e., the number of voxels V, time points T and subjects I). As such, complexity measures are obtained that are closer to the number of (independent) parameters fitted by C-ICA (i.e., the effective degrees of freedom).

Because for each component a number of scores equal to the number of timepoints, which could be 100 or more, needs to be estimated, complexity measure 2 is extended by multiplying the number of components by the number of timepoints (in complexity measure 3). Thereby, to also take the number of parameters per cluster into account, the number of clusters (in complexity measure 3) is multiplied by the total number of subjects in the clusters I_r (i.e., yielding the total number of p_{ir} values).

In complexity measure 4, a term is added where the number of components is multiplied by the number of clusters (i.e., the total number of components extracted by C-ICA). Finally, in complexity measure 5, in order to somehow take into account the number of loadings for the components, also the number of voxels is included. In this way, both the number of voxels and timepoints and the number of subjects are taken into account. Nevertheless, the complexity of a model increases more, by increasing the number of components than when increasing the number of clusters.

Table 2

Definitions of the Five Complexity Measures that could be used to quantify C-ICA's complexity

Complexity measure	Formula
Complexity 1	$Q + R$
Complexity 2	$Q \times R$
Complexity 3	$(T \times Q) + (I_r \times R)$
Complexity 4	$(T \times Q) + (I_r \times R) + (Q \times R)$
Complexity 5	$\left(T \times \frac{V}{T} \times Q\right) + (I_r \times R)$

Note. $\boldsymbol{Q}$ = number of components and $\boldsymbol{R}$ = number of clusters. The values for $\boldsymbol{T}$, $\boldsymbol{I_r}$ and $\boldsymbol{V}$ represent the number of timepoints, the number of subjects per cluster in the data and the number of voxels, respectively.

2.3.1 Model selection using CHull

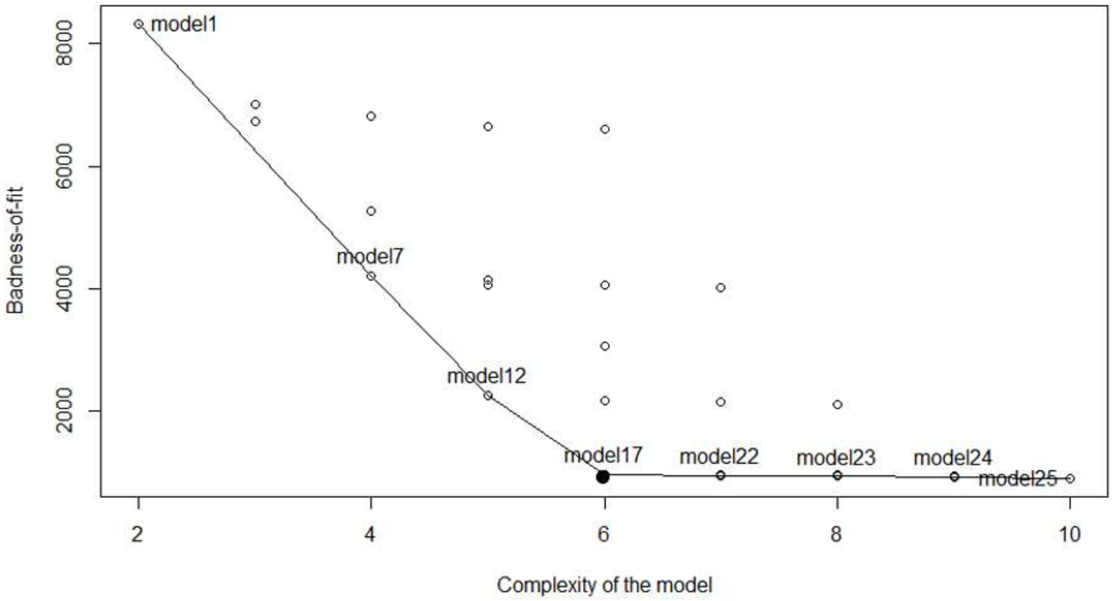
The CHull procedure selects the model that balances model goodness of fit/misfit and model complexity in some optimal way (Ceulemans & Kiers, 2006; Wilderjans et al., 2013). In general, when using a CHull analysis, a convex hull for the goodness of (mis-)fit by complexity plot is determined. After determining models that lie on the boundary of the convex hull, the optimal model is found by selecting the model that yields the largest scree-test value. Here, the scree-test value indicates how much better a solution is in comparison to a less complex one, relative to how much worse a solution is in comparison with a more complex one (Wilderjans et al., 2013). The model with the largest scree ratio is the selected model. For this model, which is on the boundary of the convex hull, it applies that when increasing the complexity there will be only a small gain in fit, while when the complexity is reduced, the fit will drop substantially.

The CHull computations consist of several steps. First, starting from a (mis)fit versus complexity plot (see, for example, Figure 1), the model with the best fit for each level of complexity is obtained. For the CHull method in this project, the goodness of fit will be the loss function (Equation 2) which followed from the C-ICA analysis from a certain dataset with a certain number of clusters R and components Q. For the complexity measure, five different options will be explored (see Table 2).

After retaining the best fitting model for each level of complexity (c_i), the CHull method automatically orders the remaining models from the simplest to the most complex one regarding the complexity measure used (Wilderjans et al., 2013). Thereafter, for every pair of adjacent models, m_i and m_j, model j will be excluded when the fit (f_j) is smaller or equal to the previous model's fit $(f_i \geq f_j)$ with $j > i$ (Wilderjans et al., 2013). Subsequently, for each trio of adjacent models - referred to as m_i, m_j and m_k - the middle model (m_j with corresponding fit f_j and complexity c_j) is excluded when $f_j \leq f_i + (c_j - c_i)\frac{(f_k - f_i)}{(c_k - c_i)}$ (Wilderjans et al., 2013). This step is repeated until there are no more models to be excluded. Now, all models which are left are located on the convex hull boundary. An example of such a convex hull, using the data as described in Section 2.1 and complexity measure 1 ($Q + R$), is illustrated in Figure 1.

Figure 1

Example Convex Hull (lower bound) plot presenting badness-of-fit (loss function value) against model complexity (measure 1 in Table 2) resulting from the C-ICA analysis on the illustrative data set.



Note. The complexity measure used in this example is complexity 1 ($Q + R$), where Q equals the number of components and R the number of clusters. Here the optimal model is model 17 ($Q = 4$ and $R = 2$).

Subsequently, for the models on the convex hull boundary, the st value is calculated using the following equation:

$$st_i = \frac{\frac{f_i - f_{i-1}}{c_i - c_{i-1}}}{\frac{f_{i+1} - f_i}{(c_{i+1} - c_i)}} . \quad (5)$$

In this equation, first, the difference in fit of a model f_i compared to the fit of the previous model (f_{i-1}) is divided by the complexity of the model c_i minus the complexity of the previous model c_{i-1}. This value is divided by the difference in fit between the current model and the next model (f_{i+1}) divided by the difference in complexity between the next model c_{i+1} and the current model. Note that both the numerator and denominator are equal to the average gain in fit per unit of complexity (however, computed when comparing different models). The final selected model is the model with the largest st value (Wilderjans et al., 2013). For the example data, using complexity measure 1, the resulting st values can be found in Table 3. Note here that for the first (most simple) and last (most complex) model, the st value is undefined and these models thus cannot be selected.

Table 3

Results CHull analysis on example dataset, indicating the st values for the models which are located on the boundary of the convex hull (the selected model is indicated in bold)

Model	Q	R	Complexity	Fit	st
1	1	1	2	8320.00	-
7	2	2	4	4571.76	1.02
12	3	2	5	2728.07	1.11
17	**4**	**2**	**6**	**1064.96**	**64.91**
22	5	2	7	1039.34	2.07
23	5	3	8	1026.96	1.07
24	5	4	9	1015.39	1.09
25	5	5	10	1004.73	-

Note. Complexity measure 1 used (Number of components Q + number of clusters R). Here, the fit-value indicates a badness-of-fit (C-ICA loss function value). For the first and last model, the st value is undefined.

As can be seen in Figure 1 and Table 3, The CHull method would in this example choose model 17 with a complexity of six, which here is the true model with 2 clusters and 4 components.

2.3.2 Model selection using Akaike Information Criterion (AIC)

This information criteria, just like CHull, tries to balance between goodness(/badness) of fit and model complexity (Akaike, 1974). This is done by using the following equation in which the negative log-likelihood $(-l_i)$ – the fit measure – is penalized by the model complexity c_i:

$$AIC = 2(-l_i) + 2 \times c_i \tag{6}$$

In this equation c_i represents the complexity of the model i, which can be defined in various ways (see Table 2). The negative log-likelihood $(-l_i)$ is derived from the C-ICA loss value (Equation 2) and is calculated as stated in Equation 3. The model with the smallest AIC value will be selected as the optimal model. As can be seen in Table 4, for the example data (when using complexity measure 1), the model with the lowest AIC value is model 25 which has 5 components and 5 clusters (which are not the correct number of clusters and components).

2.3.3 Model selection using Akaike information Criterion Corrected (AICc)

AIC is a good and unbiased estimator when the sample size is large and when the complexity of the model is relatively small (Cavanaugh, 1997). But in other situations the penalty used $(2c_i)$ can be substantially smaller than the adjustment for the bias; this exhibits a potentially high degree of negative bias, for which the AICc method tries to correct (Cavanaugh, 1997). The AICc equation is similar to the AIC, however, with AICc more complex models (with a larger complexity measure), are punished more severely:

$$AICc = AIC + \frac{2c_i(c_i+1)}{n-c_i-1} \tag{7}$$

In this equation, n is equal to the number of elements in the data (i.e., the number of elements across the $\boldsymbol{X_i}$ $(T \times V)$ multiplied by the number of all subjects) and c_i the model complexity. The model with the lowest AICc value should be selected. For the example data (see Table 4), this is, again, (the not correct) model 25 with corresponding 5 components and 5 clusters.

2.3.4 Model selection using Bayesian Information Criterion (BIC)

The BIC is a well-known method for model selection and is similar to the AIC (Neath & Cavanaugh, 2012). In contrast to the AIC method, the BIC does also take sample size into account when penalizing:

$$BIC = 2(-l_i) + \log(n) \times c_i \tag{8}$$

The negative log-likelihood $(-l_i)$ was calculated using Equations 3 and 4, just like with the AIC method. Also the complexity c_i and the number of data points n are defined as before. The model with the lowest BIC value should be selected. In general, BIC's penalty for model complexity is stronger than the one from AIC, resulting in BIC selecting more simple models than AIC. For the example data, the BIC selects the same (incorrect) model as the AIC and AICc method (see Table 4), which is a model that is too complex for this data.

2.3.5 Model selection using Kullback-Leibler Information Criterion (KIC)

The KIC method is an extended version of the AIC method. It uses the symmetric Kullback-Leibler divergence between fitted and true models (Li et al., 2007). The KIC-values were calculated using an equation which has a similar structure as the AIC equation (with symbols as defined earlier):

$$KIC = 2(-l_i) + 3 \times c_i \tag{9}$$

Similar to the previous methods, the negative log-likelihood -$l_i$was calculated using equations 3 and 4 and the model with lowest KIC value should be retained. The KIC method would, for the example data, choose the (incorrect) model with 5 components and 5 clusters (Table 4).

2.3.6 Model selection using Minimum Description Length (MDL)

This method is based on the minimum code length and also takes sample size into account (Li et al., 2007). MDL is based on the following equation (with all symbols as defined before):

$$MDL = 2(-l_i) + \frac{1}{2} \times c_i \times \log(n) \tag{10}$$

The optimal model is the model with the lowest MDL value. In the case of the sample dataset, this is a model with 5 components and 5 clusters (see Table 4).

Table 4

Models applied to the example dataset with corresponding number of components, clusters, complexity measure (measure 1 from Table 2: $Q + R$), fit measure (C-ICA loss value) and calculated values of several model order selection methods based on information theory. The selected model by each method is indicated in bold

Model	Q	R	Complexity	Badness of Fit-value	AIC	AICc	BIC	KIC	MDL
1	1	1	2	8320.00	52229.82	52229.82	52247.46	52231.82	26123.73
2	1	2	3	6894.87	42837.59	42837.59	42864.05	42840.59	21432.03
3	1	3	4	6798.96	42139.15	42139.15	42174.43	42143.15	21087.22
4	1	4	5	6729.54	41628.04	41628.04	41672.14	41633.04	20836.07
5	1	5	6	6688.78	41326.32	41326.32	41379.23	41332.32	20689.62
6	2	1	3	6664.57	41139.00	41139.00	41165.46	41142.00	20582.73
7	2	2	4	4571.76	22295.59	22295.59	22330.87	22299.59	11165.43
8	2	3	5	4427.36	20692.82	20692.82	20736.92	20697.82	10368.46
9	2	4	6	4359.67	19924.52	19924.52	19977.44	19930.52	9988.72
10	2	5	7	4262.98	18805.07	18805.08	18866.81	18812.07	9433.41
11	3	1	4	5476.77	31326.46	31326.46	31361.74	31330.46	15680.87
12	3	2	5	2728.07	-3517.64	-3517.64	-3473.54	-3512.64	-1736.77
13	3	3	6	2594.21	-6031.28	-6031.27	-5978.36	-6025.28	-2989.18
14	3	4	7	2480.52	-8269.88	-8269.88	-8208.14	-8262.88	-4104.07
15	3	5	8	2447.85	-8930.75	-8930.75	-8860.19	-8922.75	-4430.10
16	4	1	5	4336.81	19659.69	19659.69	19703.79	19664.69	9851.90
17	4	2	6	1064.96	-50548.33	-50548.33	-50495.41	-50542.33	-25247.71
18	4	3	7	1056.07	-50965.41	-50965.41	-50903.68	-50958.41	-25451.84
19	4	4	8	1047.90	-51352.07	-51352.07	-51281.51	-51344.07	-25640.76
20	4	5	9	1040.44	-51707.24	-51707.23	-51627.86	-51698.24	-25813.93
21	5	1	6	3341.77	6629.78	6629.78	6682.70	6635.78	3341.35
22	5	2	7	1039.34	-51763.96	-51763.95	-51702.22	-51756.96	-25851.11
23	5	3	8	1026.96	-52361.24	-52361.24	-52290.68	-52353.24	-26145.34
24	5	4	9	1015.39	-52925.85	-52925.85	-52846.48	-52916.85	-26423.24
25	5	5	10	1004.73	**-53451.16**	**-53451.15**	**-53362.96**	**-53441.16**	**-26681.48**

2.3.7 Model selection using DIFFIT

DIFFIT is a method that was especially developed for model order selection for the three-mode principal component analysis (3MPCA) model. In this model, the number of components for the first, second and third mode needs to be determined. As such, just like for C-ICA, the model order selection requires solving more than one selection problem (i.e., in 3MPCA the number of components can differ across modes). DIFFIT finds the optimal model order using the fit of solutions for the 3MPCA model (i.e., a least squares fit, just like in C-ICA) in combination with a complexity value s that depends on the number of components for the three modes (Timmerman & Kiers, 2000). This method, which was developed for 3MPCA, was tailored towards use for C-ICA by replacing the complexity value s by a number that is a function of R and Q (see later in Table 5).

The model selection was done in several steps, starting with the determination of the fit of all possible combinations of models with 1 to Q components and 1 to R clusters (see Table 5). The complexity values s was taken as the sum of R and Q (which corresponds with complexity measure 1 from Table 2), whereas the fit value in this case was the percentage sum of squares of the total sum of squares that was explained by the model, also called variance accounted for (VAF), calculated using this equation:

$$VAF = \frac{||X||^2 - \mathrm{L}}{||X||^2} \times 100 \tag{11}$$

In this equation, $||X||^2$ represents the sum of all squared entries of subjects' data and L is the fit-value resulting from the C-ICA analyses (Equation 2). The VAF values for the example dataset are displayed in Table 5.

Table 5

The fit of number of C-ICA models applied to the fMRI example data with several combinations of numbers of components and clusters. The complexity value is the sum of R and Q, whereas the fit is the VAF of the C-ICA solution

Q	R	$s = Q + R$	VAF (%)
1	1	2	96.56
1	2	3	97.15
1	3	4	97.19
1	4	5	97.22
1	5	6	97.24
2	1	3	97.25
2	2	4	98.11
2	3	5	98.17
2	4	6	98.20
2	5	7	98.24
3	1	4	97.74
3	2	5	98.87
3	3	6	98.93
3	4	7	98.98
3	5	8	98.99
4	1	5	98.21
4	2	6	99.56
4	3	7	99.56
4	4	8	99.57
4	5	9	99.57
5	1	6	98.62
5	2	7	99.57
5	3	8	99.58
5	4	9	99.58
5	5	10	99.59

Subsequently, the best model for each value of s, for which $Q + R = s$, was determined and selected (Timmerman & Kiers, 2000). For each selected model (displayed in Table 6), a $dif_{t(m)}$value was calculated, which is the difference in fit (%) between two consecutive models (i.e., the difference in fit between a certain model and the model after this model). Using $dif_{t(m)}$, $b_{t(m)}$ was calculated as follows: $b_{t(m)} = dif_{t(m)}/dif_{t(m+1)}$ (Timmerman & Kiers, 2000). The model, with corresponding number of clusters and number of components, with the largest $b_{t(m)}$ was selected as the optimal solution. For this model holds that, when increasing the complexity only a small gain in fit will be accomplished, while when the complexity is reduced, the fit will drop substantially. Note that this is very similar to the rationale for using st values in CHull.

In the example, see Table 6, the first model with 1 component and 1 cluster was selected as the optimal model as it has the largest $b_{t(m)}$ value. But because the first model has such a large $b_{t(m)}$ value in comparison with the other models and because this first model does not imply any clustering (R = 1), the DIFFIT method is, for this example, also conducted excluding the first model. When excluding the first model here, the model with $s = 6$ will be selected, which has 4 components and 2 clusters (and which is the true model).

Table 6

DIFFIT: selection of best fitting model per model complexity, the accompanying $\boldsymbol{dif_s}$ *value and if defined, the accompanying* $\boldsymbol{b_{t(m)}}$ *value.*

Q	R	$s = Q + R$	VAF (%)	dif_s	$b_{t(m)}$
1	1	2	96.56	96.56	111.75
2	1	3	97.25	0.68	-
2	2	4	98.11	0.86	1.14
3	2	5	98.87	0.76	1.11
4	2	6	99.56	0.69	64.91
5	2	7	99.57	0.01	2.07
5	3	8	99.58	0.01	1.07
5	4	9	99.58	0.00	1.09
5	5	10	99.59	0.00	0.00

Note: When the $\boldsymbol{dif_s}$ value for a model is lower than the $\boldsymbol{dif_s}$ for the next (more complex) model, the former (less complex) model will be excluded (i.e., $\boldsymbol{b_{t(m)}}$ is not calculated for this model). The $\boldsymbol{b_{t(m)}}$ value for the last model is equal to the second-last (included) $\boldsymbol{dif_s}$ value.

2.3.8 Overview of simultaneous methods

In the simulation study the following simultaneous model order selection techniques will be tested:

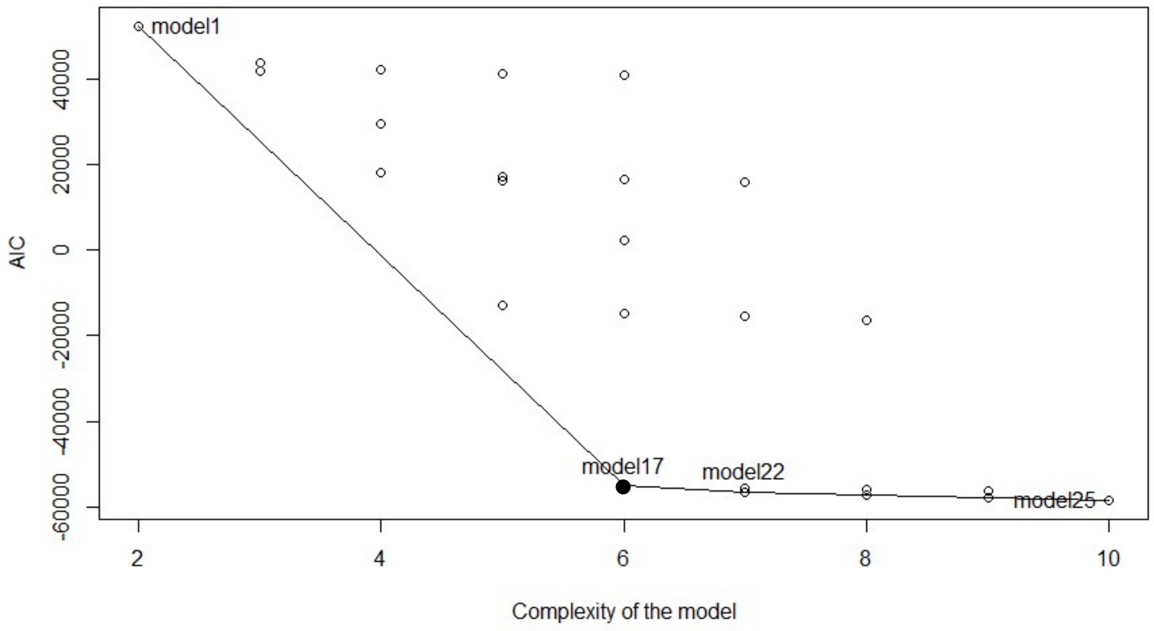
- CHull
- AIC
- AICc
- BIC
- KIC
- MDL
- DIFFIT

All these methods (except for DIFFIT) make use of a complexity measure/penalty which can be freely chosen. Therefore, these methods will be performed five times, using the five different complexity measures as listed in Table 2. For the information theoretic measures (AIC, AICc, BIC, KIC and MDL) this includes that for every complexity measure the model with the lowest corresponding value will be selected as optimal model. Note that, although possible in theory, the definition of s as $Q + R$ for DIFFIT will not be varied using the complexity measures from Table 2. The reason is that the original DIFFIT model is only presented with that particular definition of s.

Besides this variation in complexity measure, for selection methods AIC, AICc, BIC, KIC and MDL, the optimal number of components and clusters will also be determined by using CHull (instead of simply choosing the lowest value). The reason for this is that a small pilot study showed that the information theoretic measures often select the most complex model (what is not desirable). To solve this issue the combination of CHull with the information theoretic measures is done by performing a CHull analysis, as explained in Section 2.3.1, but now using the resulting values (for example AIC values) as goodness of fit values and adopting the complexity measures from Table 2. In Figure 2 below, one example is illustrated using the CHull on AIC results. Here, the badness of fit, used in CHull, is defined by the AIC-values as stated in Table 4 and the complexity is complexity measure 1 as defined in Table 2.

Figure 2

Convex hull (lower bound) resulting from the CHull analysis on the example dataset, using the AIC-value as fit-measure



Note. The complexity measure used in this example is complexity 1 ($Q + R$), where Q equals the number of components and R the number of clusters. Here the optimal model is model 17 ($Q = 4$ and $R = 2$).

After retaining the convex hull as illustrated above, the st-values are calculated in the same way as for the regular CHull method, explained earlier. In this example, the CHull method would, for AIC, choose a model with a complexity of six, which, here is a model with 2 clusters and 4 components (see model 17 in Table 1). This procedure will be repeated five times using the five different complexity measures listed in Table 2. For the remaining methods, this procedure is similar, but they use their own corresponding values instead of the AIC value (Table 4).

2.4 Sequential methods

In the sequential methods, both selection problems (i.e., identifying the number of components Q and clusters R) are handled sequentially. For C-ICA there are two possibilities: (1) first selecting Q and next (using the optimal Q) R and (2) first selecting R and next (using the optimal R) Q. Both possibilities will be used in the simulation study to see whether the order of solving the selection problems leads to different models being selected.

In the first sequential methods (Section 2.4.1), scree ratios are used to determine R and Q sequentially. Next, in Section 2.4.2, two methods are discussed that can be used to select the number of components Q irrespective of the number of clusters R. In Section 2.4.3, two methods are discussed to select the number of clusters R irrespective of the number of components Q. In the final section (2.4.4), the sequential methods that will be evaluated in the simulation study are listed.

2.4.1 Model selection using **scree ratios**

Sequential model selection using VAF and scree ratios is done in several steps. Here, we first will select R and next Q. The first step involves computing the total sum of squares ($\|X\|^2$) across the data of all subjects. With this value, the VAF can be computed for each combination of number of clusters and number of components. The computation of the VAF value is done in the same way as for DIFFIT (see Equation 11 and Table 5). Secondly, the scree ratio for each number of clusters R ($r = 2, \dots, R-1$), fixing for a certain number of components Q, was computed (and this was computed for each possible value of Q: q = 1, …, Q):

$$Scree\ ratio\ R\ |(Q = q) = \frac{VAF_{(r-1,q)} - VAF_{(r,q)}}{VAF_{(r,q)} - VAF_{(r+1,q)}} \qquad (12)$$

In this equation, $VAF_{(r,q)}$ represents the VAF for a certain model with r clusters and q components, as stated in Table 1. This ratio is large when deleting a cluster implies a serious decrease in VAF and adding a cluster only improves VAF a little. After obtaining all the scree ratios for each number of clusters R, the average scree ratio (averaged across $q = 1, \dots, Q$) is computed and the value of R associated with the largest average is selected as the optimal number of clusters. Note that the lowest and largest value of R cannot be selected as no scree values can be computed for these values (i.e., there is no smaller or larger R value to compare with). The computation for the example data set is illustrated in Table 7. Here, the optimal number of clusters R is equal to 2, with an average scree ratio of 120.78. Sequentially, using this optimal number of clusters, the number of components was determined using Equation 13, fixing for the optimal R:

$$Scree\ ratio\ Q|R = \frac{VAF_{Q-1} - VAF_{Q}}{VAF_{Q} - VAF_{Q+1}} \qquad (13)$$

The largest number resulting from this formula, indicates the optimal Q. Applying this procedure to the small data example results in the selection of the correct model with 2 clusters and 4 components (see Table 7).

Table 7

Result of the sequential model selection procedure (selecting R first) applied to the example data

	$R = 2$	$R = 3$	$R = 4$	Optimal Q
Q = 1	14.86	1.38	1.70	-
Q = 2	14.49	2.13	0.70	1.26
Q = 3	20.53	1.18	3.48	1.10
Q = 4	368.08	1.09	1.10	**64.91**
Q = 5	185.94	1.07	1.09	-
Average	**120.78**	1.37	1.61	

Note. Computed scree-ratios $\mathrm{sr}_{\mathrm{r|q}}$ from step 1 of the procedure are displayed in the columns $R = 2$, $R = 3$ and $R = 4$ for all values of Q (rows) and averaged across Q (bottom row). The computed scree ratios $\mathrm{sr}_{\mathrm{q|R}}$ -conditional on the optimal R- from step 2 of the procedure are presented in the last column. Scree-ratios are not defined for $R_{min} = 1$ and $R_{max} = 5$. Similarly, for the second step, the scree ratios are not defined for $Q_{min} = 1$ and $Q_{max} = 5$. The average value of the optimal R and Q are indicated in bold.

To see whether the order matters, that is first computing R or Q, also a technique will be tested in which Q is determined first and next R. To determine Q, the scree ratios are computed as:

$$Scree\ ratio\ Q\ |(R = r) = \frac{VAF_{(r,q-1)} - VAF_{(r,q)}}{VAF_{(r,q)} - VAF_{(r,q+1)}} \tag{14}$$

and averaged across all R values. Next, given an optimal Q, the optimal value of R is determined by calculating

$$Scree\ ratio\ R|Q = \frac{VAF_{(Q,R-1)} - VAF_{(Q,R)}}{VAF_{(Q,R)} - VAF_{(Q,R+1)}} \tag{15}$$

and calculating the R for which this ratio is maximal. The calculations for this are illustrated in Table 8, resulting in a (correct) solution with 2 clusters and 4 components.

Table 8

Result of the sequential model selection procedure in reversed order (selecting Q first) applied to the example data

	$Q = 2$	$Q = 3$	$Q = 4$	Optimal R
R = 1	1.39	1.04	1.15	**-**
R = 2	1.26	1.11	64.91	**368.08**
R = 3	1.29	1.19	52.83	1.09
R = 4	1.26	1.31	44.07	1.10
R = 5	1.34	1.29	39.42	-
Average	1.31	1.19	**40.47**	

Note. Computed scree-ratios $sr_{q|r}$ from step 1 of the procedure are displayed in de columns $Q = 2$, $Q = 3$ and $Q = 4$ for all values of R (rows) and averaged across R (bottom row). The computed scree ratios $sr_{r|Q}$ -conditional on the optimal Q- from step 2 of the procedure are presented in the last column. Scree-ratios are not defined for $Q_{min} = 1$ and $Q_{max} = 5$. Similarly, for the second step, the scree ratios are not defined for $R_{min} = 1$ and $R_{max} = 5$. The average value of the optimal R and Q are indicated in bold.

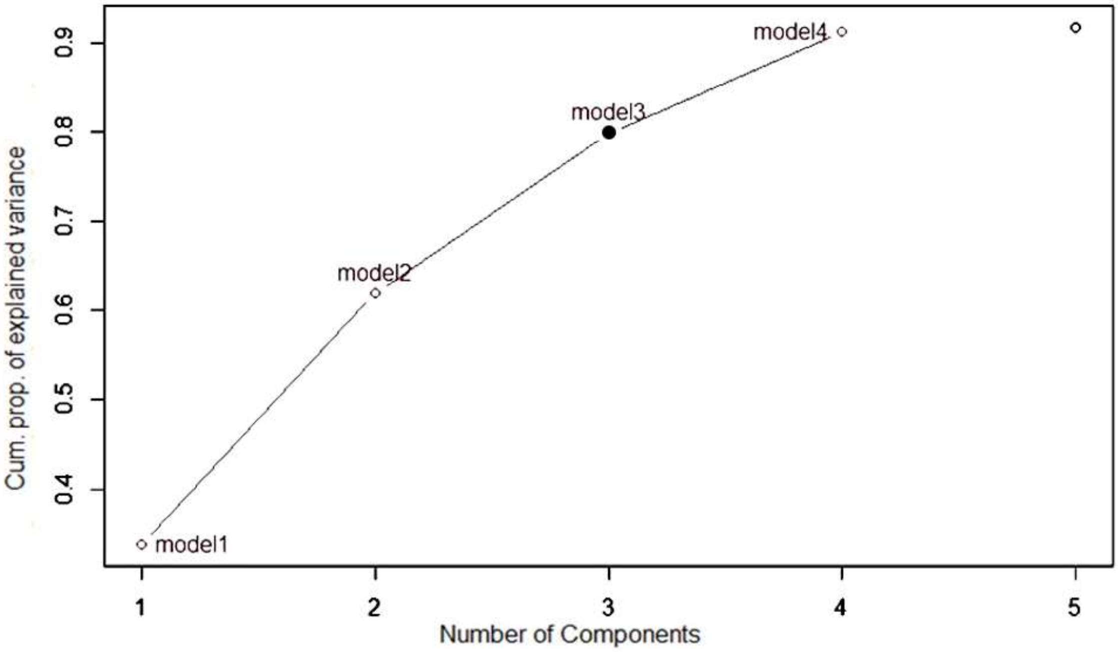
2.4.2 Selecting the number of components Q

2.4.2.1 Selecting *Q* using Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique (James et al., 2017). This method puts correlated variables together into a few uncorrelated principal components, to reduce the number of variables (Richardson, 2009). In this thesis, for each subject's dataset, a PCA analysis was performed. To select, for each subject, the optimal number of PCA components, a CHull analysis is performed. So, for each subject, a CHull analysis was done with the cumulative proportion of explained variance as fit measure and the number of retained components as complexity measure, resulting in an optimal number of components for each subject. An example of a CHull analysis using PCA results to one (of the ten) subject's data is illustrated in Figure 3.

Figure 3

CHull analysis on PCA data of a single subject from the example dataset



Note. Convex hull resulting from the CHull analysis using an upper bound, with cumulative proportion of explained variance of the PCA analysis (fit value) and number of components (complexity value) conducted on the first subject's data of the example dataset. Here the selected optimal model is model 3 (where $Q = 3$).

The corresponding *st*-values for the models on the convex hull are shown in Table 9. For this specific subject, a number of 2 components should be selected, according to the CHull analysis.

Table 9

Results CHull analysis using PCA on the data of the first subject of the example dataset, indicating the st values for the models which are located on the convex hull

Model	Complexity	Fit	st
1	1	0.33	-
2	2	0.62	1.56
3	3	0.80	**1.59**
4	4	0.91	-

Note. Here, the fit-value indicates the cumulative proportion of explained variance which results from the PCA analysis on the subject's dataset. The complexity refers to the number of components. The optimal model is indicated in bold. Note that CHull cannot select the simplest and most complex model.

The distribution of the selected number of components Q across subjects in the example dataset is illustrated in the top row of Table 10. To select a single optimal number of components for the whole data set, the mode of the distribution of optimal Q across subjects was retained as the final number of components Q. In the example data, this results in 2 components.

Table 10

Frequencies of the selected number of components per subject for PCA and pPCA

	Q = 2	Q = 3	Q = 4	Q = 5	Optimal Q
PCA	**6**	4	0	0	2
pPCA	0	0	**10**	0	4

Note. For the example data, the PCA and pPCA is performed on ten subjects' data. In the case of multiple modes, the smallest value of Q (of the modes), is selected as optimal number of components Q.

2.4.2.2 Selecting Q using Probabilistic Principal Component Analysis (pPCA)

This method is a version of PCA based on a Gaussian latent variable model (Tipping & Bishop, 1999). A pPCA analysis can be performed on each subject's dataset to compute the optimal number of components per subject (as was done with PCA and CHull in Section 2.4.2.1). This can be done by calculating the following log-likelihood (L), which is optimized by pPCA, for each number of components Q for each subject (Tipping & Bishop, 1999):

$$L = -\frac{N}{2}\{d\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-\mathbf{1}}\mathbf{S})\} \tag{16}$$

where

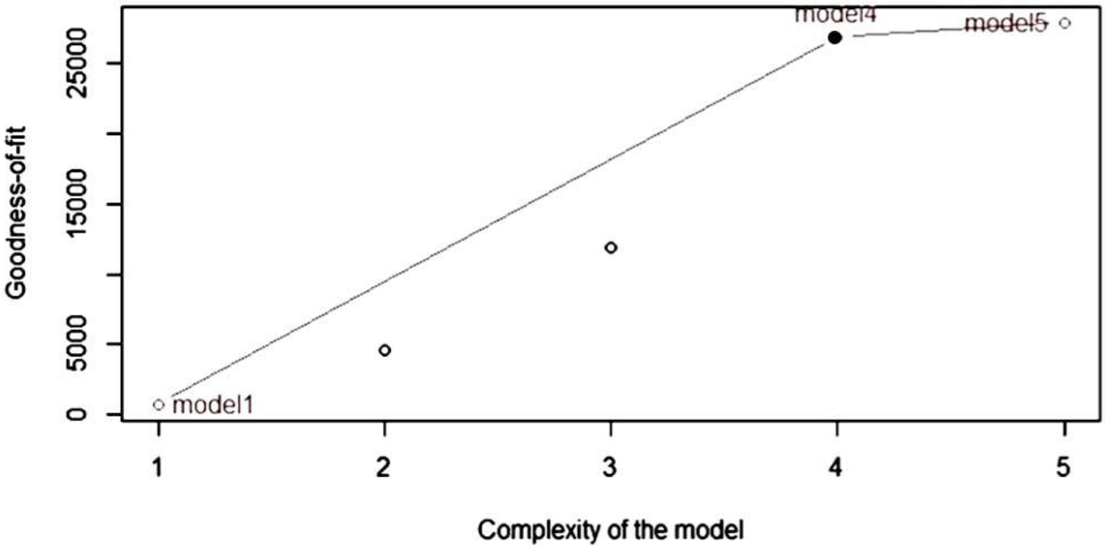
$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{t_n} - \boldsymbol{\mu})(\boldsymbol{t_n} - \boldsymbol{\mu})^{\mathrm{T}} \tag{17}$$

and $\mathbf{C} = \mathbf{W}\mathbf{W}^{\boldsymbol{t}} + \sigma^2\mathbf{I}$, with $\boldsymbol{W}$ being the Lambda matrix (with component loadings) and σ^2 the noise variance and both parameters resulting from the pPCA analysis. In this formula, N represents the number of rows and d the number of columns in a subject's dataset. So, for the example data, the number of rows is 100 (timepoints) and the number of columns 50 (voxels).

To select the optimal number of components for this method, for each subject's dataset, the loglikelihood (Equation 16) and its corresponding number of pPCA components was used in a CHull analysis. This resulted in an optimal number of components for every subject. An example of one subject's dataset is illustrated in Figure 4.

Figure 4

CHull analysis on probabilistic PCA data of a single subject from the example dataset



Note. Convex hull resulting from the CHull analysis using an upper bound, with the log-likelihood of the pPCA analysis (fit value) and number of components (complexity value) conducted on the first subject's data of the example dataset. Here the selected optimal model is model 4 (where $Q = 4$).

The corresponding st-values are for the models which lie on the boundary of the convex hull are shown in Table 11. According to these results, for this subject, a number of 4 underlying components was selected as optimal.

Table 11

Results CHull analysis using pPCA on the data of the first subject of the example dataset, indicating the st values for the models which are located on the convex hull

Model	Complexity	Fit	st
1	1	678.37	-
4	4	26919.88	**8.73**
5	5	27922.04	**-**

Note. Here, the fit-value indicates the loglikelihood-value which results from the pPCA analysis on the subject's dataset. The complexity refers to the number of components. The optimal model is indicated in bold. Note that CHull cannot select the simplest and most complex model.

Again, the mode of the distribution of optimal Q values across subjects can be taken as the optimal Q for the full multi-subject data set. For the empirical example, as can be seen in the bottom row of Table 10, the optimal Q equals 4.

2.4.3 Selecting the number of clusters R

2.4.3.1 Selecting R by using CHull results of K-means clustering on vectorized data

K-means is a method which divides a (two-way) dataset (i.e., objects by variables) in K distinct and non-overlapping clusters (James et al., 2017). The main idea of this method is to cluster the objects in such a way that the within-cluster variation is as small as possible and the between-cluster variation as large as possible. In order to apply K-means to a multi-subject rs-fMRI data set (which has a three-way structure: time by voxel by subject) each subjects' data matrix ($T \times V$) was vectorized. This results in N vectors of size $T \times V$, which can be stored in a $N \times TV$ matrix, in which each row represents the data of a single subject. Next, K-means was applied to this vectorized and concatenated data with the number of clusters ranging from 1 to R. CHull was used to identify the optimal number of clusters, with the model (mis)fit (i.e., sum of squared errors for each number of clusters) as fit measure and R as complexity measure.

An example of a CHull analysis using K-means results and the corresponding *st*-values for the models on the convex hull are shown in Table 12. According to these results, a solution with 4 clusters should be selected.

Table 12

Results CHull analysis using K-means data on the example dataset, indicating the st values for the models which are located on the convex hull

Model	Complexity	Fit	st
1	1	220076.60	-
2	2	191118.90	1.07
3	3	163929.80	1.04
4	4	137715.50	**1.08**
5	5	113359.10	-

Note. Here, the fit-value indicates the sum of squared errors which results from the K-means analysis on the example dataset. The complexity refers to the number of clusters. The selected model is indicated in bold.

2.4.3.2 Selecting R by using the Gap Statistic results of K-means clustering on vectorized data

The Gap statistic is a method for determining the number of clusters present in a data set. This method uses the pooled within-cluster sum of squares around the cluster means, W_k:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{18}$$

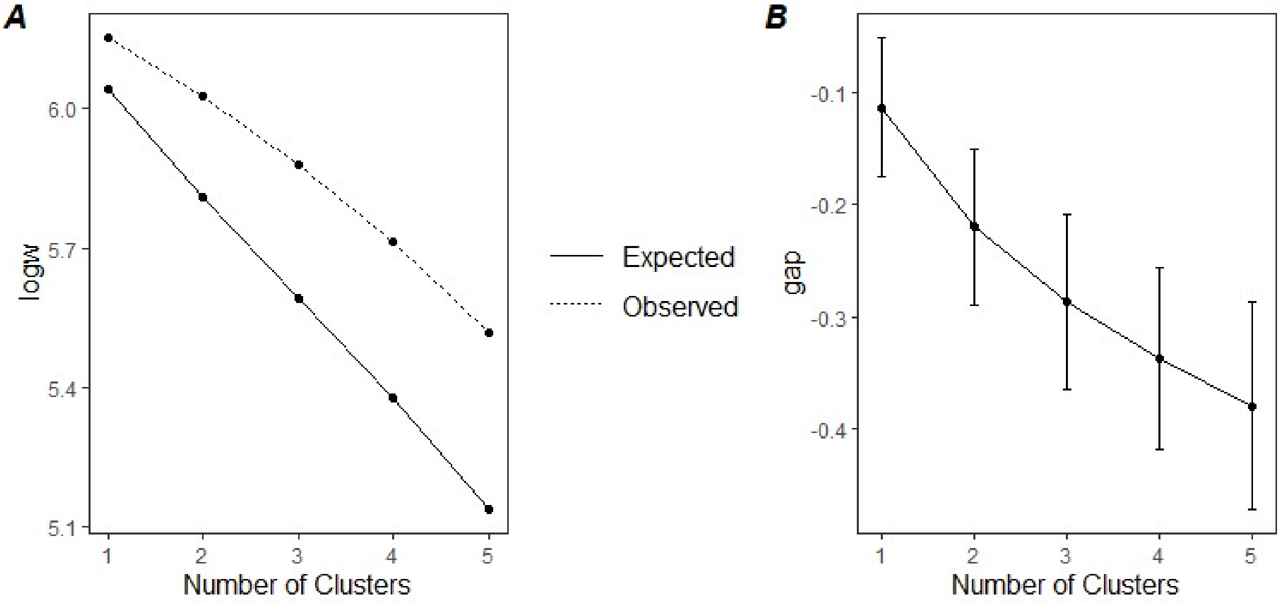
Here, D_r indicates the sum of the pairwise squared Euclidean distances for points within a cluster r, with sample size n_r of a cluster (Tibshirani et al., 2001). The clusters are a result of a K-means clustering algorithm, applied to the concatenated $N \times TV$ matrix (as explained in Section 2.4.3.1). To select the number of clusters, $\log(W_k)$ is compared to an expected appropriate null reference distribution of the data with a sample of size n (E_n^*), this is defined as (Tibshirani et al., 2001):

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k) \tag{19}$$

The log of the expected and observed pooled within-cluster sum of squares, for the sample data is illustrated in Figure 5a. The difference between those two, also known as the Gap, is shown in Figure 5b. The model with the largest Gap statistic is the optimal model. In the example, a model with 1 cluster is selected.

Figure 5

Gap statistic analysis on example dataset



Note. (Panel A) The log of the expected and observed pooled within-cluster sum of squares, for the sample data. (Panel B) The difference between those two (with confidence bounds), also known as the Gap. The model with the largest Gap value should be selected.

2.4.4 Overview of sequential model order techniques

For the methods which only estimate one of both outcomes, which are pPCA, PCA, Gap statistic and K-means clustering, all possible combinations will be tested. Each combination contains one cluster selection method and one component selection method. This implies that in the simulation study the following sequential model order selection techniques will be tested:

- Scree ratios for VAF-values with first determining Q and next R
- Scree ratios for VAF-values with first determining R and next Q
- PCA with CHull for determining Q and Gap statistic for determining R
- pPCA with CHull for determining Q and Gap statistic for determining R
- PCA with CHull for determining Q and K-means with CHull for determining R
- pPCA with CHull for determining Q and K-means with CHull for determining R

III. Simulation Study

3.1. Problem

A difficult task before performing a C-ICA analysis is specifying the number of components and clusters which should be estimated. Generally, there is no information about how many components and clusters are optimal for a given data set at hand. To address this problem, a simulation study will be conducted in which several model order selection methods will be tested and compared to each other and it will be determined which method is most efficient. Both simultaneous and sequential selection methods will be compared (see Section 2). In addition, it will be investigated whether the performance of the model order selection methods depends on data characteristics such as the amount of noise in the fMRI data.

Regarding previous research, the CHull methods seems to work well for model selection in the context of applying dimension reduction and a combination of dimension reduction and clustering on complex and big datasets (Rossbroich et al., 2022). For, for example, mixtures of factor analysers CHull even outperformed AIC and BIC (Bulteel et al., 2013). It is therefore expected that CHull will work well for fMRI data. For information theoretical methods like AIC and BIC it is expected that they will only work well when a good measure of model complexity is used. As it is unclear how the model complexity for C-ICA should be quantified optimally (i.e., effective degrees of freedom), some model complexity measures may work better than others. Thereby it is expected that, when comparing simultaneous and sequential methods, the sequential methods in comparison to simultaneous methods may increase estimation errors. The reason for this is that the selection of a wrong number of components (or clusters) in an earlier step may negatively influence the determination of the number of clusters (or components) in a later step of the analysis.

Finally, characteristics of the data, such as amount of noise in the data and the number of components underlying the data, could be influential on the results of model selection methods. Earlier research showed that large noise levels in fMRI data could exaggerate estimates of the number of components in ICA (Majeed & Avison, 2014). It is therefore expected that, for C-ICA, similar effects would be perceived. Larger levels of noise in the data would result in worse estimates of number of components and clusters in C-ICA.

3.2 Simulation Study

3.2.1 Design

In this simulation study the number of subjects per cluster was fixed at 20. Thereby, the number of voxels (2000) and length of a time course (100) were also fixed to make the design not to complex. The remaining three factors which will be systematically varied are:

1. The true number of components Q, with four levels: 2, 5, 25 and 50;
2. The true number of clusters R, with two levels: 2 and 4;
3. The percentage of noise in the data ε, with three levels: 10%, 40% and 70%.

3.2.2 Procedure

To generate data under the C-ICA model, first, a binary partition matrix $\boldsymbol{P}$ $(I \times R)$ which determines the cluster to which each subject belongs (with R clusters), is generated in such a way that each cluster exactly has 20 subjects. Further, cluster specific sources matrices $\boldsymbol{S^r}$

(with Q components) were simulated by using the R function *icasamp* from the ica package (Helwig, 2018). This procedure resulted in no overlap between the $\boldsymbol{S^r}$'s as the average pairwise RV coefficient between the $\boldsymbol{S^r}$'s equals 0. Next, subject specific time courses $\boldsymbol{A_i}$ were generated by randomly and independently drawing values from a uniform distribution $U(-2,2)$. The p_{ir} (values in P), $\boldsymbol{S^r}$ and $\boldsymbol{A_i}$ were combined into true data $\boldsymbol{T_i}$ $(i = 1, \ldots, I)$ for each subject through the C-ICA model formulation $\boldsymbol{T_i} = \sum_{r=1}^{R} p_{ir}\boldsymbol{A_i}\boldsymbol{S^r}$. Finally, noise $\boldsymbol{E_i}$ was generated by drawing random numbers form $\mathcal{N}(0,1)$. This Gaussian noise was scaled in order to ensure the required amount of noise in the data ε being 10%, 40% and 70%, respectively. In particular, $\boldsymbol{E_i}$ was scaled by equalling the sum of squared entries (SSQ) of the noise matrices to the corresponding SSQ of the (noiseless) data blocks $\boldsymbol{T_i}$ and next a portion of $\boldsymbol{E_i}$ was added to $\boldsymbol{T_i}$ (i.e., $\boldsymbol{X_i} = \boldsymbol{T_i} + \omega\boldsymbol{E_i}$) in such a way that the required noise percentage was obtained (i.e., by choosing an appropriate value for ω). This resulted in data matrices $\boldsymbol{X_i}$ for each subject $(i = 1, \ldots, I)$.

For each cell in the design, which contains all factor combinations, 10 replication data sets were generated. This results in a total of 4 (number of components) × 2 (number of clusters) × 3 (percentage noise) × 10 (replications) = 240 generated datasets.

C-ICA (with 10 multistarts) was performed to each generated data set, with the number of components ranging from 1 up to 75 (1 to 5, from 5, with steps of 5 to 75: 1, 2, 3, 4, 5, 10, …, 75) and the number of clusters going from 1 to 6. For each data set separately, the different model order selection techniques and the variations on those techniques (i.e., 62 in total; see Table 13) were applied to the obtained C-ICA models and the optimal number of components and clusters retained by each technique was recorded.

Important to note is that, due to time constraints, for the selection methods in which a selection method for the number of components (pPCA and PCA) was combined with a selection method for the number of clusters (K-means and Gap statistic), only the first replication of each cell in the design was performed. As a consequence, these (sequential) methods were applied on only 24 datasets.

Table 13

Overview of all model order selection methods and their variations used in this thesis

Method	Variations of method	Total number of analyses per method
Simultaneous		
CHull	• Using 5 different complexity measures	5
AIC	• Using 5 different complexity measures (selecting the solution with the lowest value) • Using CHull with 5 different complexity measures	10
AICc	• Using 5 different complexity measures (selecting the solution with the lowest value) • Using CHull with 5 different complexity measures	10
BIC	• Using 5 different complexity measures (selecting the solution with the lowest value) • Using CHull with 5 different complexity measures	10
KIC	• Using 5 different complexity measures (selecting the solution with the lowest value) • Using CHull with 5 different complexity measures	10
MDL	• Using 5 different complexity measures (selecting the solution with the lowest value) • Using CHull with 5 different complexity measures	10
DIFFIT	• None	1
Sequential		
Scree ratio's/VAF	• Scree ratios for VAF-values with first determining Q and next R • Scree ratios for VAF-values with first determining R and next Q	2
PCA-Gap	• PCA with CHull for determining Q and Gap statistic for determining R	1
pPCA-Gap	• pPCA with CHull for determining Q and Gap statistic for determining R	1
PCA-K means	• PCA with CHull for determining Q and K-means with CHull for determining R	1
pPCA- K means	• pPCA with CHull for determining Q and K-means with CHull for determining R	1
	Total:	62

Several R packages were used in the data analysis. The code for the CHull method was derived from Wilderjans, et al. (2013). The PCA and K-means clustering analysis were performed using the '*stats*' R-package by R Core Team (2021). The Gap statistics was calculated using the '*cluster*' R-package by Maechler et al. (2021). C-ICA and the remaining model order selection methods were performed using custom made code (Durieux, 2021).

To evaluate how the different methods perform, an estimation error is calculated by comparing the estimated number of components Q_{est} and the estimated number of clusters R_{est} to the optimal/true number of components Q_{true} and optimal/true number of clusters R_{true}:

$$estimation\ error = 0.5 * \left| \frac{(Q_{true} - Q_{est})}{(Q_{max} - Q_{min})} \right| + \left| \frac{(R_{true} - R_{est})}{(R_{max} - R_{min})} \right| \qquad (17)$$

Here, the maximum number of components Q_{max} and maximum number of clusters R_{max}, which are specified in the simulation study, are equal to 75 and 6, respectively. The minimum number of components Q_{min} and minimum number of clusters R_{min} are both equal to one. The estimation error ranges from 0 to 1. The lower the estimation error the better the method performs. Besides estimation error, also an accuracy measure was computed which equals one when the correct model with the correct number of components AND clusters was retained and zero otherwise.

IV Results

In this section, the performance results of the various model selection methods are presented. First, the results for the simultaneous model selection methods are presented, followed by the results for the sequential methods. Next, the effects of data characteristics (i.e., number of clusters, number of components and amount of noise) will be evaluated. Subsequently, the influence of the different complexity measures (See Table 2) is discussed. Finally, the best simultaneous- and best sequential method will be compared and evaluated in more detail.

4.1 Performance of methods

4.1.1 Simultaneous methods

In Table 14, the mean estimation error overall and for each level of each manipulated data characteristic separately is presented for each simultaneous model order selection method. For CHull, AIC, AICc, BIC, KIC, MDL and the combinations of CHull and the information theoretic measures, the prediction errors of the methods are averaged over all the five complexity measures (C1-C5; see Table 2; results per complexity measure are presented in Appendix A1-A5, see further).

The table shows that overall (bottom row), CHull is the best performing simultaneous method with a small mean estimation error of 0.12. The methods which are second best in overall estimating the model order are the information theoretic measures in combination with CHull: AIC, AICc and KIC (all 0.13), with BIC and MDL following closely (0.14). Using the information theoretic measures without CHull selection (i.e., taking the solution with the lowest value on AIC, BIC, etc.) yields bad results (mean estimation error of 0.66). DIFFIT performs in between the other methods with a mean estimation error of 0.33. Again, note that the mean estimation error for CHull, the information theoretic measures and the combination of both are an average of these errors over all five complexity measures. Differences in mean estimation errors per complexity measure as well as the percentage of totally correct selected models (per complexity measure and method), can be found in Appendix A1-A6. The patterns regarding best performing methods, stated in these tables, are similar to the pattern found in Table 14.

For almost all levels of each data characteristic, the CHull method (or a combination of CHull with an information theoretic measure) was the most effective in terms of minimizing estimation error, except for datasets where the number of components was 25. In this specific case, CHull in combination with an information criterion performed slightly better.

When looking at specific cases of analyses where (the worst performing) information theoretic measures (AIC, BIC, etc.) are used, it is striking that these methods often choose the most complex model (which results in a large estimation error). When using those information theoretic measures in combination with CHull, those estimation errors decreased substantially (as can be seen in Table 14). Interestingly, the information theoretic measures without CHull perform better when the true number of components underlying the data increases, whereas the reverse is true for these methods in combination with CHull. Finally, the information theoretic measures without CHull are insensitive to the amount of noise in the data, whereas these methods with CHull, as expected, perform worse with increasing noise levels.

Table 14

*Mean estimation errors (*with *SD) for the simultaneous model order selection methods, computed overall and per level of the manipulated factors*

	Level	CHull	AIC	CHull	AICc	CHull	BIC	CHull	KIC	CHull	MDL	CHull	DIFFIT	Overall
				AIC		AICc		BIC		KIC		MDL		
Number of	2	0.05*	0.79	0.05*	0.79	0.05*	0.74	0.06	0.79	0.05*	0.74	0.06	0.19	0.37
components		(0.03)	(0.10)	(0.03)	(0.10)	(0.03)	(0.05)	(0.03)	(0.10)	(0.03)	(0.05)	(0.03)	(0.12)	(0.35)
	5	0.05*	0.77	0.05*	0.77	0.05*	0.73	0.05*	0.77	0.05*	0.73	0.05*	0.23	0.36
		(0.03)	(0.1)	(0.02)	(0.1)	(0.02)	(0.05)	(0.02)	(0.10)	(0.02)	(0.05)	(0.02)	(0.10)	(0.34)
	25	0.18	0.63	0.17*	0.63	0.17*	0.59	0.17*	0.63	0.17*	0.59	0.17*	0.36	0.37
		(0.07)	(0.11)	(0.06)	(0.11)	(0.06)	(0.09)	(0.07)	(0.11)	(0.06)	(0.09)	(0.07)	(0.10)	(0.23)
	50	0.18*	0.46	0.26	0.46	0.26	0.44	0.27	0.46	0.26	0.44	0.27	0.53	0.36
		(0.11)	(0.10)	(0.13)	(0.10)	(0.13)	(0.08)	(0.14)	(0.10)	(0.13)	(0.08)	(0.14)	(0.10)	(0.16)
Number of	2	0.14*	0.76	0.15	0.76	0.15	0.69	0.15	0.76	0.15	0.69	0.15	0.22	0.40
clusters		(0.13)	(0.14)	(0.15)	(0.14)	(0.15)	(0.11)	(0.15)	(0.14)	(0.15)	(0.11)	(0.15)	(0.14)	(0.32)
	4	0.10*	0.57	0.12	0.57	0.12	0.56	0.12	0.57	0.12	0.56	0.12	0.43	0.33
		(0.03)	(0.13)	(0.06)	(0.13)	(0.06)	(0.13)	(0.06)	(0.13)	(0.06)	(0.13)	(0.06)	(0.13)	(0.24)
Amount of noise	.1	0.09*	0.67	0.10	0.67	0.10	0.63	0.10	0.67	0.10	0.63	0.10	0.32	0.34
		(0.05)	(0.17)	(0.08)	(0.17)	(0.08)	(0.14)	(0.08)	(0.17)	(0.08)	(0.14)	(0.08)	(0.19)	(0.29)
	.4	0.12*	0.67	*0.12*	0.67	*0.12*	0.63	0.13	0.67	*0.12*	0.63	0.13	0.33	0.36
		(0.10)	(0.16)	(0.10)	(0.16)	(0.10)	(0.14)	(0.11)	(0.16)	(0.10)	(0.14)	(0.11)	(0.16)	(0.28)
	.7	0.14*	0.66	0.18	0.66	0.18	0.62	0.18	0.66	0.18	0.62	0.18	0.33	0.38
		(0.11)	(0.17)	(0.14)	(0.17)	(0.14)	(0.15)	(0.14)	(0.17)	(0.14)	(0.15)	(0.14)	(0.16)	(0.27)
Overall		**0.12**	0.66	0.13	0.66	0.13	0.62	0.14	0.66	0.13	0.62	0.14	0.33	0.36
		(0.09)	(0.17)	(0.11)	(0.17)	(0.11)	(0.14)	(0.12)	(0.17)	(0.11)	(0.14)	(0.12)	(0.17)	(0.28)

Note. The method with the lowest overall mean estimation error is indicated in bold; the best method(s) per data characteristic is indicated with a *. For CHull, the information theoretic measures (AIC, AICc, BIC, KIC and MDL) and the combinations of CHull and the information theoretic measures, the estimation errors are averaged over the five different complexity measures. For performance results per method per complexity measure, see Appendix A1-A5.

4.1.2 Sequential methods

The mean estimation error for the sequential methods, both overall and per level of each manipulated factor separately, are displayed in Table 15. The two methods which both use scree ratios on VAF data (but in reversed order) seem to perform very well and this especially when first the number of clusters and next the number of components is selected (mean estimation error of 0.002 versus 0.040 vice versa). The combined use of component- and cluster selection methods did not yield such accurate results. Of those methods, the best combination was using PCA with K-means clustering (0.19), whereas PCA with the Gap statistic performed a bit worse (0.27). Using probabilistic PCA (pPCA) in combination with K-means clustering and Gap statistic gave similar results (mean estimation error of 0.20 and 0.27, respectively). Note that for these combined methods only 24 datasets were used (instead of 240 like the other methods) due to time constraints.

The ordering of the sequential methods in terms of performance is quite consistent across all levels of all design factors, with the VAF scree method with estimating the number of clusters first being the best and pPCA with the Gap statistic the worst performing method. Similar results can be found when looking at the percentage correctly estimated models (see Appendix A7). The VAF scree method with first selecting the number of clusters and subsequently the number of components selects the correct model in 97.50% of the cases.

Table 15

*Mean estimation errors (*with *SD) for the sequential model order selection methods, computed overall and per level of the manipulated design factors*

	Level	VAF scree Components-clusters	VAF scree Clusters-components	PCA-Gap	PCA-K means	pPCA-Gap	pPCA-K means	Overall
Number of components	2	0.00* (0.00)	0.00* (0.00)	0.20 (0.11)	0.14 (0.08)	0.20 (0.11)	0.14 (0.08)	0.03 (0.07)
	5	0.12 (0.12)	0.00* (0.01)	0.20 (0.11)	0.12 (0.10)	0.21 (0.11)	0.13 (0.10)	0.08 (0.11)
	25	0.03 (0.11)	0.01* (0.04)	0.18 (0.11)	0.10 (0.10)	0.26 (0.15)	0.18 (0.15)	0.04 (0.11)
	50	0.01 (0.04)	0.00* (0.00)	0.5 (0.11)	0.42 (0.10)	0.44 (0.22)	0.36 (0.22)	0.07 (0.17)
Number of clusters	2	0.00 (0.03)	0.00* (0.00)	0.18 (0.14)	0.14 (0.15)	0.17 (0.13)	0.13 (0.16)	0.03 (0.08)
	4	0.08 (0.12)	0.00* (0.03)	0.37 (0.14)	0.25 (0.15)	0.39 (0.14)	0.27 (0.14)	0.08 (0.14)
Amount of noise	.1	0.03 (0.08)	0.00* (0.00)	0.23 (0.14)	0.16 (0.13)	0.17 (0.10)	0.10 (0.11)	0.03 (0.09)
	.4	0.03 (0.08)	0.00* (0.02)	0.29 (0.18)	0.19 (0.20)	0.29 (0.18)	0.19 (0.20)	0.05 (0.12)
	.7	0.06 (0.12)	0.00* (0.03)	0.28 (0.18)	0.22 (0.15)	0.33 (0.18)	0.27 (0.14)	0.07 (0.14)
Overall		0.04 (0.10)	**0.00 (0.02)**	0.27 (0.17)	0.19 (0.16)	0.27 (0.17)	0.20 (0.16)	0.05 (0.12)

Note. The method with the lowest overall mean estimation error is indicated in bold; the best method(s) per data characteristic is indicated with a *.

4.2 Influence of data characteristics on performance

4.2.1 Mean estimation error per data characteristic

Simultaneous methods. According to Table 14, the estimation error for each level of the number of components seems to be quite equal (mean estimation error of = .37, .36, .37 and .36 for Q= 2, 5, 25 and 50, respectively). The mean estimation error seems to increase when the data has a smaller number of clusters (mean estimation error = .33 and .40 for R = 4 and 2, respectively) and a larger amount of noise (mean estimation error = .34, .36 and .38 for amount of noise = .1, .4 and .7, respectively).

Sequential methods. As can be seen in the last column in Table 15, where the mean estimation error for each level of data characteristic is displayed, the overall estimation error is mostly impaired (1) for datasets containing 5 underlying components (mean estimation error = .08) compared to the other levels (est. error = .03, .04, .07 for Q = 2, 25, 50, respectively), (2) for data sets with larger number of clusters (mean estimation error = .08) compared to lower number of clusters (mean estimation error = .03), and (3) for data with larger amounts of noise (mean estimation error = .03, .05 and .07 for Error = .1, .4 and .7, respectively).

4.2.2 Mixed Factorial ANOVA

To further investigate the effects of the dataset characteristics on performance, a mixed factorial ANOVA with the estimation error as dependent variable and the design factors (true number of underlying clusters, true number of underlying components and amount of noise; between factor) and method (within factor with 14 levels) as independent variables was performed including all possible main and interaction effects. For the selection methods which make use of a complexity measure, the averaged mean estimation error (over the five complexity measures) is used (similar as in Table 14). Due to the fact that for the combinations of the selection methods PCA and pPCA with Gap statistic and K-means clustering only 24 datasets were used (instead of 240 datasets for the other methods), these methods were excluded from the ANOVA analysis. Lastly, in the ANOVA analysis the generalized eta-squared (η_G^2) is used as an indicator for effect size for all main and interaction effects (Bakeman, 2005).

From Table 16, which shows the ANOVA results of the methods which used all 240 datasets, it appears that all main and interaction effects are significant at .001 level, even after a Greenhouse-Geiser correction for violation of sphericity, except for the interaction between the number of clusters and amount of noise (*p* = .240). Only discussing effects with a substantial effect size (i.e., larger than 0.50), it seems that the strongest effect on mean estimation error is the choice of model selection method (η_G^2 = .99), which suggests strong differences in performance between the different methods. Overall, CHull or a combination of information theoretic measures with CHull clearly outperformed the other simultaneous methods (see Table 14). Thereby, the scree-ratios on VAF data (in both orders) performed strikingly better than CHull. The number of clusters a dataset has, seems to have a strong effect on the mean estimation error (η_G^2 = .68) as well. Overall, when the true number of clusters of a dataset equals 2 it seems to result in larger estimation errors (mean estimation error = .34) than for data where the true number of clusters is larger (mean estimation error = .29). Further, the amount of noise in the data also seems to have some -although not very large- influence on performance (η_G^2 = .35), with performance deteriorating when the data contain more noise (see Tables 14 and 15). Finally, the number of components influences performance to a small -negligible- degree (η_G^2 = .07).

Table 16

ANOVA table presenting the effects of data characteristics and method on mean prediction errors for the model selection methods according to a mixed factorial ANOVA analysis. Substantial effects (generalized eta squared (η_G^2) larger than .50) are indicated in bold

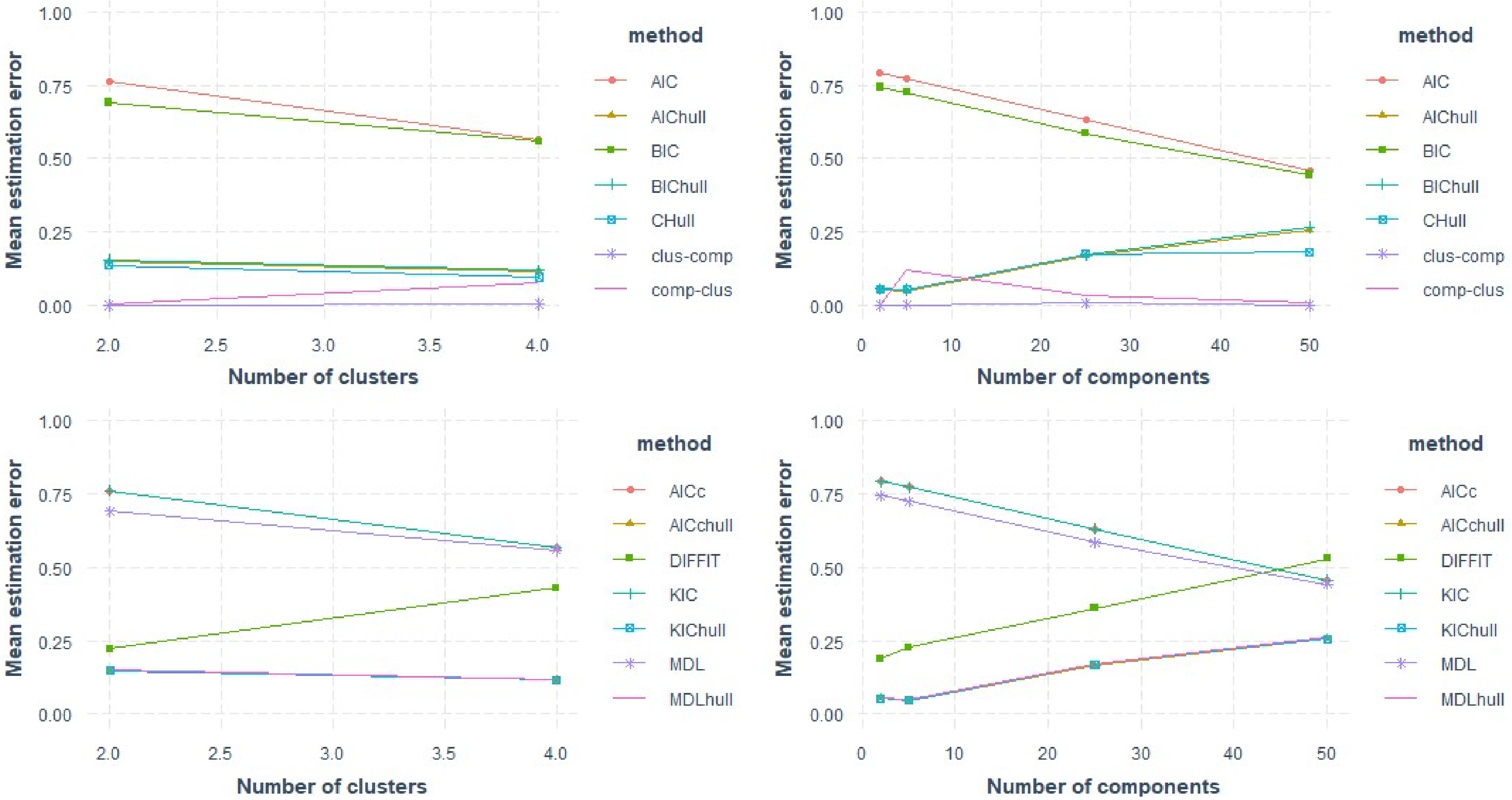
Effect	*DF*	*F*-value	*p-value*	η_G^2
Components	3	29.103	< .001	0.07
Clusters	**1**	**2393.079**	**< .001**	**0.68**
Noise	2	297.394	< .001	0.35
Method	**13**	**48903.327**	**< .001**	**0.99**
Components × Clusters	**3**	**652.319**	**< .001**	**0.64**
Components × Noise	6	83.989	< .001	0.31
Clusters × Noise	2	1.435	0.24	0.00
Components × Method	**39**	**2533.759**	**< .001**	**0.97**
Clusters × Method	**13**	**2150.638**	**< .001**	**0.89**
Noise × Method	26	103.716	< .001	0.44
Components × Clusters × Noise	6	17.527	< .001	0.09
Components × Clusters × Method	**39**	**185.206**	**< .001**	**0.68**
Components × Noise × Method	78	39.949	< .001	0.47
Clusters × Noise × Method	26	11.721	< .001	0.08
Components × Clusters × Error × Method	78	15.009	< .001	0.25

The main effects of method and the number of clusters are qualified by an interaction between them (η_G^2 = .89), an interaction between the method and the number of components (η_G^2 = .97) and an interaction between the number of clusters and the number of components (η_G^2 = .64). To interpret these interactions, the two interactions with the largest effects (i.e., interaction of number of clusters with method and number of components with method) are visualized in Figure 6. From this figure, one can see that for the good performing methods (e.g., CHull, information theoretic measures in combination with CHull and scree ratios on VAF data in both orders) estimation error is not influenced much by the number of underlying clusters; for the best performing methods (scree ratios on VAF), performance decreases a little with increasing number of clusters, with the opposite being true for the other good performing methods. For bad performing methods, however, performance increases with increasing the numbers of clusters (except for DIFFIT for which the reverse holds).

Regarding the other interaction, it seems that the estimation error increases a little for good performing methods like CHull and the information theoretic methods in combination with CHull. For the best performing method (scree ratio on VAF, R first), this increase is very small, where for the second-best performing methods (scree ratio on VAF, reversed order), the pattern is less univocal (i.e., worst performance for an intermediate number of components). For the worst performing methods (i.e., information theoretic measures), however, the mean estimation error decreases when the underlying data structure becomes more complex (i.e., more clusters and components). This could indicate that these methods often select the most complex models; this indication corresponds with the findings in our pilot study (see Section 2.3.8 and a discussion hereof in Section 5.2).

Figure 6

Two-way interactions between number of clusters and method (left) and number of components and method (right). The 14 model order selection methods are divided across two panels (7 methods in the top panels and the other 7 methods in the bottom panels)

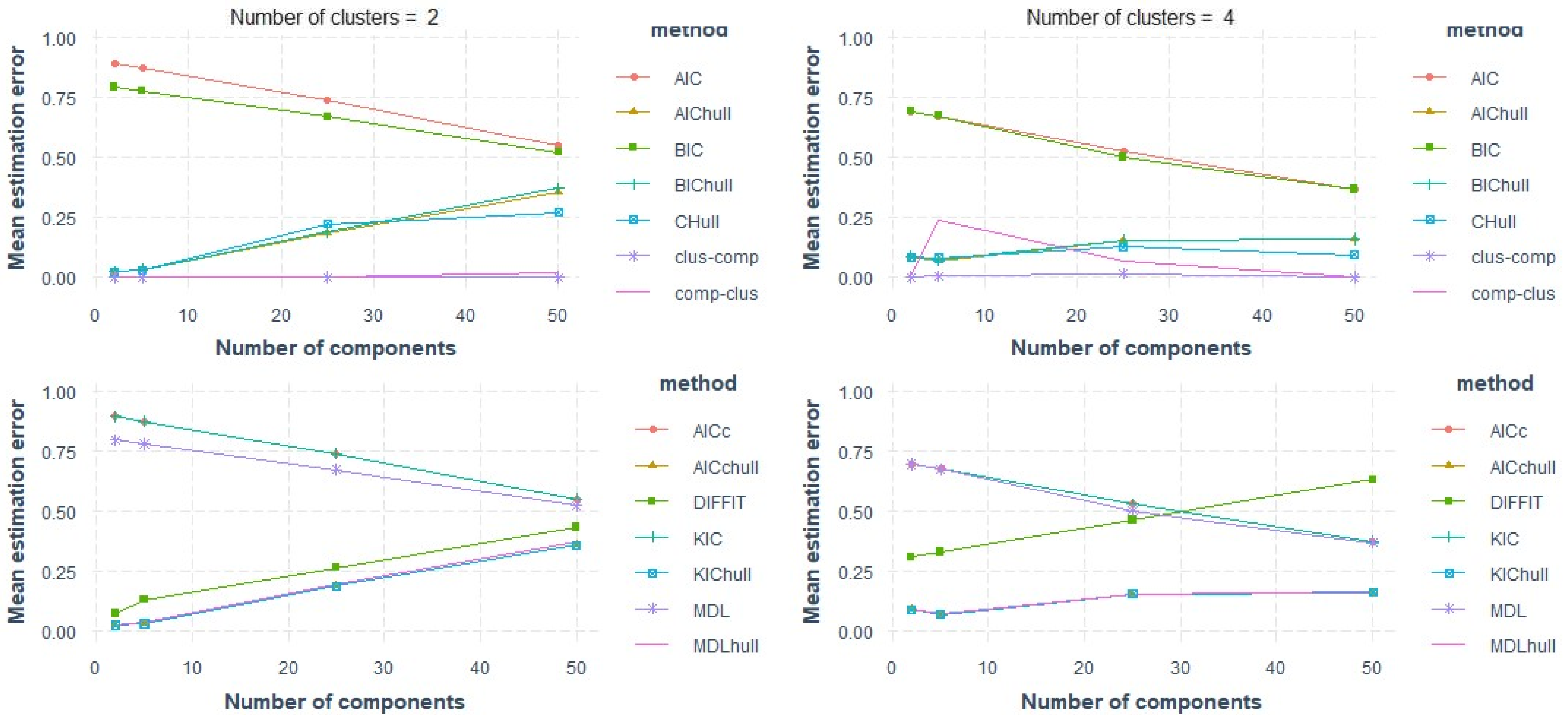


Note. Here clus-comp and comp-clus are referring to the method which uses scree ratios on VAF data; selecting the number of clusters first (clus-comp) and selecting the number of components first (comp-clus).

Lastly, all previously discussed main and interaction effects are qualified by a three-way interaction between method, the number of clusters and the number of components ($\eta_G^2 =$ .68). This complex interaction is presented in Figure 7. From this figure, it appears that for the best performing method (scree ratio on VAF) there are no big differences in pattern for the two-way interaction between methods and the number of components when comparing data with small versus large number of clusters. For the second best performing method (scree ratio on VAF, reversed order), there is a little peak in estimation error when the number of components is 5 and the number of clusters is high (compared to when the number of clusters is low). For the methods which uses CHull (or a combination with CHull) it appears that the effect of the number of components (i.e., larger prediction error with increasing Q) is more pronounced (i.e., larger increase) for data sets with two underlying clusters than for data sets with four clusters. Finally, for the (worst performing) information theoretic measures the patterns in both lower- and large number of clusters are comparable. However, for these methods in general the lower the number of clusters the higher the estimation errors are, with the opposite being true for DIFFIT.

Figure 7

Three-way interaction between number of components, method and number of clusters (left panels: number of clusters is low, right panels: number of clusters is high). The 14 model order selection methods are divided across two panels (7 methods in the top panels and the other 7 methods in the bottom panels).



Note. Here clus-comp and comp-clus are referring to the method which uses scree ratios on VAF data; selecting the number of clusters first (clus-comp) and selecting the number of components first (comp-clus).

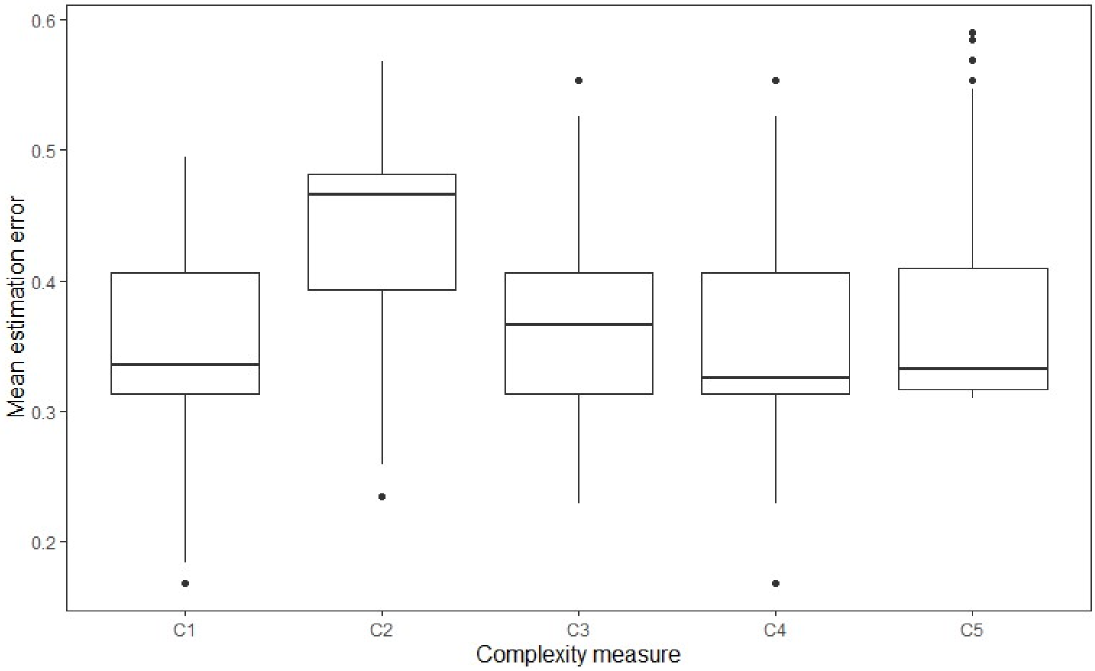
4.3 Influence of the measure for model complexity

All the simultaneous methods (except DIFFIT) make use of a certain penalty for model complexity (complexity measure), of which the computation of this measure can be freely chosen. In this thesis, these methods are used with all the five different complexity measures as listed in Table 2.

Figure 8 shows the estimation error (in boxplots) for the different complexity measures overall (i.e., across all design factors). From this figure, one can see that complexity measure 2 in general results in larger estimation errors and therefore performs the worst of all five complexity measures. The best performing measure, on average, seems C4, closely followed by C5 and C1. There is, however, quite some spread around the mean, which makes it difficult to differentiate between C1 and C3-C5 in terms of performance. C5, for example, has quite some outliers with a bad performance and is, similar to C3 and C4, skewed towards these worse performing scores, whereas C1 is more skewed towards better performing scores. It is a bit unexpected that C1 and C5 performs at the same level as C1 is the smallest considered complexity value and C5 the largest one. C2, which has an intermediate complexity value, performs clearly worse than C1 and C5, whereas C3 and C4, which are also intermediate complexity values, perform at the same level as C1 and C5. One can conclude that the effect of complexity measure is not univocal.

Figure 8

Boxplot of estimation errors for each of the five complexity measure (C1-C5) computed across all data sets

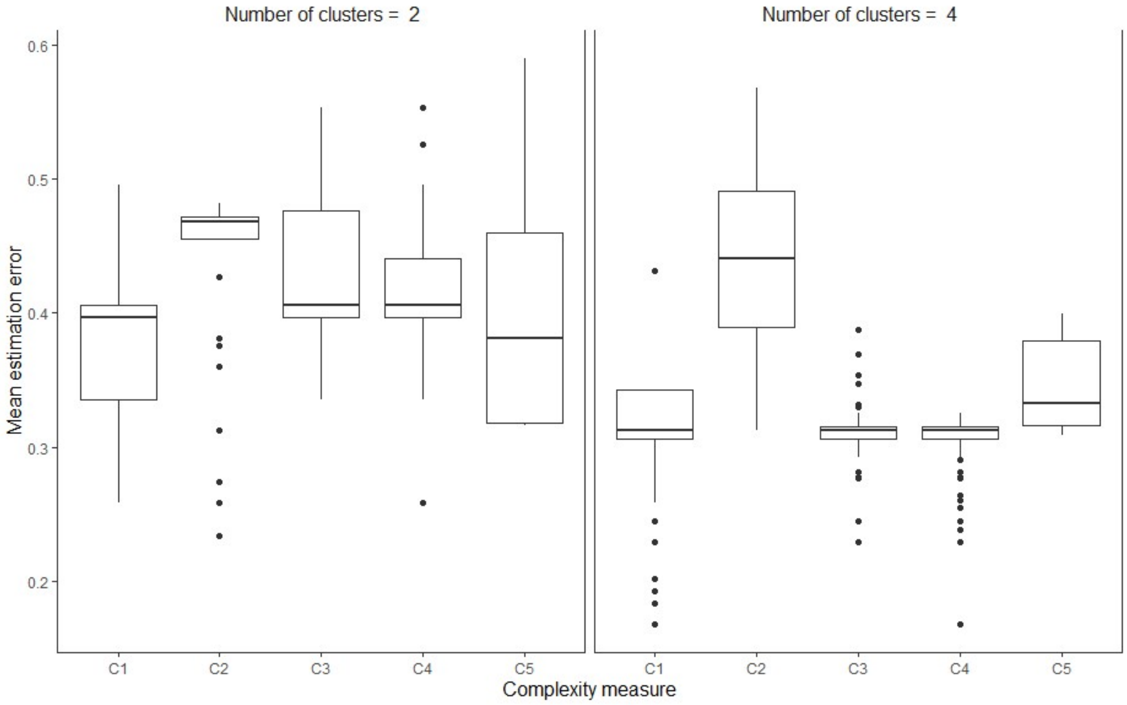


Note. Here C1-C5 refers to the complexity measures 1 - 5 used (see Table 2). C1 = $Q + R$, C2 = $Q \times R$, C3 = $(T \times Q) + (I_r \times R)$, C4 =$(T \times Q) + (I_r \times R) + (Q \times R)$ and C5 = $\left(T \times \frac{V}{T} \times Q\right) + (I_r \times R)$; Q equals the number of components, R number of clusters, T number of timepoints, I_r the total number of subjects per cluster in the data and V the number of voxels.

In Figure 9, the same boxplots as in Figure 8 are shown but now distinctions are made between datasets where the true number of clusters equals 2 (left panel) or equals 4 (right panel). This plot indicates that when using complexity measure 2 (C2 = $Q \times R$), the estimation errors are larger compared to the other complexity measures, with this effect being more pronounced when the number of clusters is larger (in this case 4 clusters). There are no striking differences between the other four complexity measures. Complexity measure 1 (C1) seems to work slightly better than C3-C5 when the number of clusters is equal to 2; this difference in performance is smaller when the true number of clusters equals 4. Remarkably, the variance in estimation error decreases with increasing number of clusters and this especially for C3-C5.

Figure 9

Boxplot of estimation errors for each of the five complexity measures (C1-C5) for two (left panel) and four (right panel) clusters

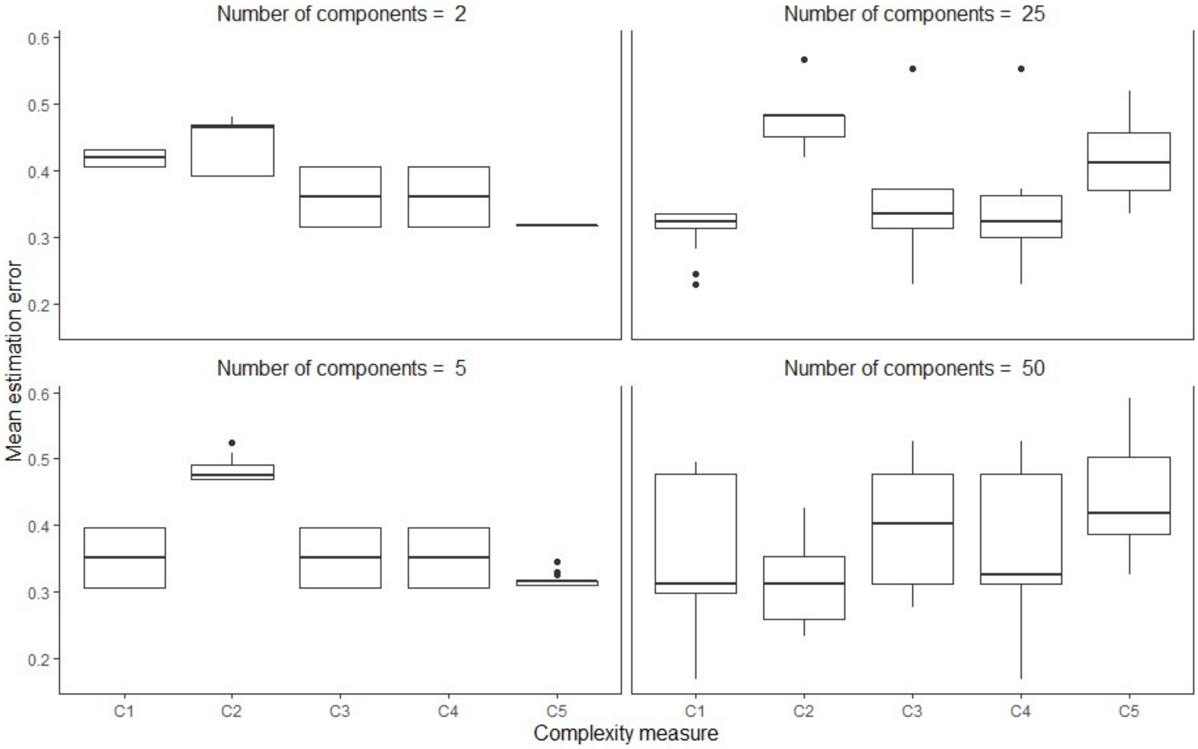


Note. Here C1-C5 refers to the complexity measures 1 - 5 used (see Table 2). C1 = $Q + R$, C2 = $Q \times R$, C3 = $(T \times Q) + (I_r \times R)$, C4 =$(T \times Q) + (I_r \times R) + (Q \times R)$ and C5 = $\left(T \times \frac{V}{T} \times Q\right) + (I_r \times R)$; Q equals the number of components, R number of clusters, T number of timepoints, I_r the total number of subjects per cluster in the data and V the number of voxels.

Figure 10 illustrates boxplots of the estimation errors for each complexity measure where distinctions are made between the true number of components. It is remarkable that, again, complexity measure 2 (C2) does not seem to work well compared to the other four. However, when the number of components is relatively large (in this case; 50), C2 actually performs a bit better than the other four methods (Figure 9; bottom right panel). Measures C3 and C4 perform very similar in all cases. Compared to C3-C5, measure C1 performs worse for 2 components (although C1 has less variability there), at the same level for 5 and 25 components (with less variability for C1 with 25 components) and a bit better for 50 components. Overall, the plots suggests that the safest choice for complexity measure would be C1 ($Q + R$), except when the true number of expected components is very low. Interestingly, the variation in estimation errors for all complexity measures is larger for data sets with 50 components than for data sets with a smaller number of components.

Figure 10

Boxplot of estimation errors for each of the five complexity measure (C1-C5) for 2 (upper left), 5 (bottom left), 25 (upper right) and 50 (bottom right) components

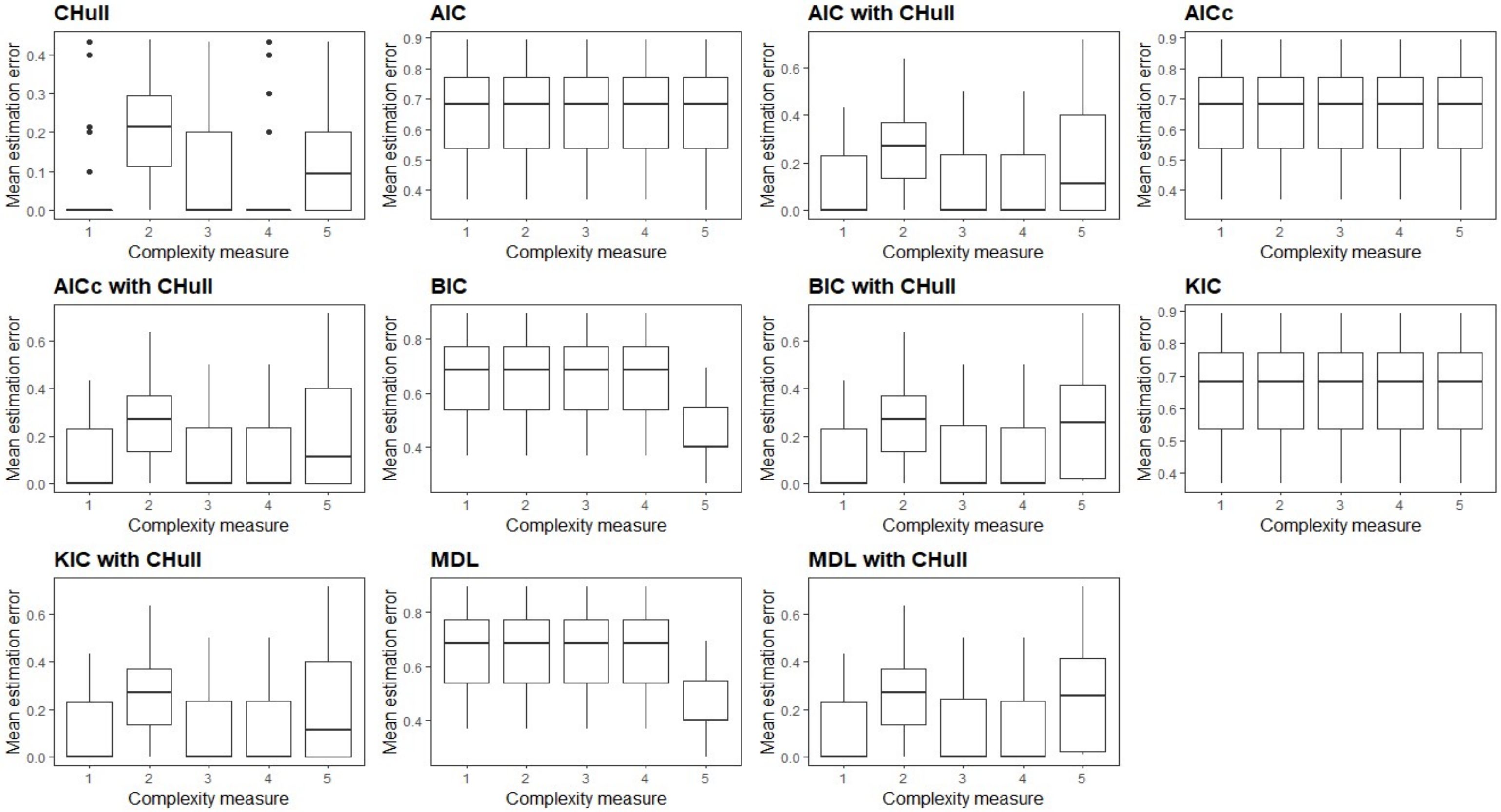


Note. Here C1-C5 refers to the complexity measures 1 - 5 used (see Table 2). C1 = $Q + R$, C2 = $Q \times R$, C3 = $(T \times Q) + (I_r \times R)$, C4 =$(T \times Q) + (I_r \times R) + (Q \times R)$ and C5 = $\left(T \times \frac{V}{T} \times Q\right) + (I_r \times R)$; Q equals the number of components, R number of clusters, T number of timepoints, I_r the total number of subjects per cluster in the data and V the number of voxels.

Figure 11 shows boxplots of estimation errors for each complexity measure split out for each model order selection method. It appears that for the methods that (in some way) uses CHull, complexity measure 1 performs best, very closely followed by C3 and C4 and that complexity measure 2 performs the worst. For the method that uses CHull solely very good results are obtained with C1 and C4, whereas C3 performs a bit worse. For the information theoretic measures, where the model with the lowest (AIC, AICc, etc.) value was selected, all complexity measures performed equally bad. Except for BIC and MDL, here complexity measure 5 seems to perform a little bit better than the other four complexity measures (but still at an unsatisfactory level). Note that the BIC penalizes complexity quite strongly. It is remarkably that especially the largest complexity values here yields the best results (see discussion Section 5.2).

Figure 11

Boxplot of mean estimation errors for each of the five complexity measures (C1-C5) for each method



Note. Here C1-C5 refers to the complexity measures 1 - 5 used (see Table 2). C1 = $Q + R$, C2 = $Q \times R$, C3 = $(T \times Q) + (I_r \times R)$, C4 =$(T \times Q) + (I_r \times R) + (Q \times R)$ and C5 = $\left(T \times \frac{V}{T} \times Q\right) + (I_r \times R)$; Q equals the number of components, R number of clusters, T number of timepoints, I_r the total number of subjects per cluster in the data and V the number of voxels.

4.4 Comparing the best performing methods

Regarding the simultaneous methods, overall, CHull seems to perform best (Table 14), except when the number of components was equal to 25. But, in this case, the difference in mean estimation error of CHull compared to the best performing method (i.e., information theoretic measures combined with CHull) was very small. Because CHull needs a complexity measure, the effect of the choice for this measure on performance is investigated further, in order to identify the most optimal procedure for CHull. According to the results from Section 4.3, complexity measure 1 -closely followed by complexity measure 4- seems to be the best option. This is in line with the results in the first five columns of the first row in Table 17, which presents the mean estimation error for CHull in combination with each of the five complexity measures. However, when looking at the accuracy (i.e., % data sets for which the model selection method identifies the true number of clusters AND components), CHull using complexity measure 4 seems to perform slightly better than when using complexity measure 1 (see bottom row of Table 17). In general, the best sequential method which uses scree ratios on VAF data, where first the number of clusters are estimated and subsequently the number of components, clearly outperforms the CHull procedure (and all other simultaneous and sequential methods) both in terms of mean estimation error (0.004) and accuracy (97.5%).

Table 17

Comparing the best performing simultaneous method (CHull using the five different complexity measures) with the best sequential model selection method (VAF using scree ratios with first estimating the number of clusters) in terms of estimation error and accuracy (% correct)

	CHull C1	CHull C2	CHull C3	CHull C4	CHull C5	Sequential
Mean est. error (*SD*)	0.06 (*0.13*)	0.19 (*0.14*)	0.12 (*0.16*)	0.07 (*0.15*)	0.14 (*0.15*)	0.00 (*0.02*)
% Correct	77.50	23.33	63.75	78.33	45.83	97.50

Note. Here C1-C5 refers to the complexity measures 1 - 5 used (see Table 2). C1 = $Q\ +\ R$, C2 = $Q \times R,$ C3 = $(T \times Q) + (I_r \times R)$, C4 =$(T \times Q) + (I_r \times R) + (Q \times R)$ and C5 = $\left(T \times \frac{V}{T} \times Q\right) + (I_r \times R);$ Q equals the number of components, R number of clusters, T number of timepoints, I_r the total number of subjects per cluster in the data and V the number of voxels.

V. Discussion

5.1 Summary

The current thesis investigated several methods to address the model order selection problem for C-ICA. To find out which method is most accurate in estimating the optimal number of underlying clusters and components for C-ICA, a simulation study was performed. A strong effect of choice of method on estimation error was found, indicating large differences in performance between the methods. Overall, CHull outperformed all the other simultaneous methods, a finding which is in line with our expectations and previous research (for example, Bulteel et al., 2013). CHull performed better than other methods, except when the number of underlying components was equal to twenty-five. In this specific case, a combination of CHull with information theoretic measures like AIC or AICc with CHull outperformed CHull, although the differences in estimation error was minimal.

The safest choice of complexity measure for the simultaneous methods seems to be $Q + R$. This was also the case when specifically looking at the estimation error of CHull, which was the best performing simultaneous method. However, when looking at the percentage of totally correctly estimated models, using $(T \times Q) + (I_r \times R) + (Q \times R)$ as complexity measure seemed to work slightly better for CHull. Note that for the other simultaneous methods, in most cases, this latter complexity measure did only perform a little bit worse than $Q + R$.

Regarding the comparison of simultaneous and sequential methods, it was expected that the sequential methods would increase the estimation error. The reason for this was that the selection of a wrong number of components (or clusters) in an earlier step may negatively influence the determination of the number of clusters (or components) in a later step of the analysis. Nevertheless, overall, results suggest that a sequential method based on scree ratios applied to the Variance Accounted For (VAF) values, actually outperformed (the simultaneous method) CHull. Moreover, first determining the number of clusters and subsequently the number of components performed better than a scree ratio based method in which the optimal number of components was identified before the optimal number of clusters. These results are quite consistent across the manipulated data characteristic (i.e., amount of noise, true number of components and clusters). The other sequential methods (i.e., combinations of PCA and pPCA with the Gap statistic and K-means) did not perform very well.

The manipulated data characteristics also influenced how well the different selection methods estimated the number of clusters and components. Focusing on main effects, for the simultaneous methods, the number of underlying clusters was influencing the estimation errors the most; on average, the higher the number of clusters the more accurate the simultaneous methods were. Also for the sequential methods, the number of clusters seemed to have the largest effect. However, the influence of the number of underlying clusters was different for both types of methods. Indeed, more accurate estimations were obtained when the number of clusters was low. Regarding the number of components, for the simultaneous methods no striking differences were found, whereas for the sequential methods the best performance was encountered when the number of components was very small. Lastly, for all methods (i.e., simultaneous and sequential ones), overall, the higher the amount of noise was the worse the methods performed; this is similar to the conclusion that Majeed & Avison (2014) have drawn in their study. With respect to the interactions between the factors, a sizeable three-way interaction between method, the number of components and the number of clusters was observed. It appears that the best performing method (Scree ratio on VAF, R

first) is stable and well performing in all conditions. For, the second-best methods- CHull (or a combination with CHull)- the performance is stable when the number of clusters is large but estimation error increases when the number of components is large and number of clusters small.

5.2 Limitations and issues for further research

Results suggests that measures from information theory like AIC, AICc, BIC, KIC and MDL but also DIFFIT and combinations of sequential methods (PCA and pPCA combined with K-means and Gap-statistic) do not perform very well in estimating the true number of clusters and components for C-ICA. When looking at these results individually, it is remarkable that many of these methods are selecting a model that is extreme in complexity. Note that this pattern can also be seen in some of the examples provided in Section 2. In other words, these methods do often select a model which contains the minimum or maximum number of clusters and components. For example, in Table 6 of section 2.3.7, it can be seen that DIFFIT selects the first model (which contains 1 cluster and 1 component). However, when this model would not have been included, DIFFIT would have chosen the correct model. Similar results were found in a small pilot study. So, when using these methods in further research, it is advisable to consider running these algorithms while excluding the most extreme model or models. This will result in a “second best” model that possibly may return better performance results (i.e., lower estimation errors). Moreover, as most of these methods rely on a particular choice of complexity measure, the complexity measures considered in this thesis may not be optimal for these methods. Using more extreme complexity measures may increase the estimation performance of these methods. Rossbroich et al. (2022), for example, show that in the context of ADPROCLUS, which is also a clustering method like C-ICA, the performance of information theoretic measures can be improved by changing the induced penalty, and this even above the performance level of CHull. An indication in this regard in this thesis is the better performance of the largest complexity measure (C5) for the information theoretic measures (without combination with CHull). These results suggest that using an even larger complexity measure may increase the performance of the information theoretic measures. For future research, it is therefore recommended to study what would be an appropriate measure to quantify the complexity of a C-ICA model and to also consider larger complexity values than C5.

Another important limitation in this study relates to the implementation of the sequential methods which only estimate either the number of clusters or the number of components. Due to time constraints, not all 240 datasets were analyzed. A decision has been made to only apply these methods on the first iteration of each cell of the simulation design. This implies that only results are presented for 24 generated datasets, one from each cell of the simulation design. As such, variability in the performance results is not considered for sequential methods, which makes the results for this type of methods less stable. In future research, the stability and generalizability of the performance results for the sequential methods should be investigated.

Finally, regarding the order of selection for the sequential scree-ratios on VAF data method, results show that first selecting the number of clusters and subsequently the number of components yields the lowest estimation error. Important to take into account regarding these results is the number of possible options for selecting these optimal number of components and clusters. The current thesis made use of a maximum of 6 clusters and 75 components during the analysis. A consequence of this is that the chance of selecting the number of clusters correctly is higher -because there are less options to choose from- than selecting the correct number of components. Concluding, when using this method in further research, it should still be investigated whether first estimating the number of clusters is always performing better than first estimating the number of components. From a substantive point of view, however, it can be expected that the number of components almost always will be larger than the number of clusters, except maybe for a big data set in which many subjects should be clustered and it can be expected that many clusters are needed for this. Therefore, for an applied user, it is quite logic to try more different values for the number of components than for the number of clusters. For fMRI data, for example, the optimal number of components is often located between 20 and 70, whereas 10 clusters is already a large number of clusters.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379-384.

Barkhof, F., Haller, S., & Rombouts, S. A. R. B. (2014). Resting-state functional MR imaging: a new window to the brain. *Radiology*, *272*(1), 29-49. doi:10.1148/radiol.14132388

Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360 (*1457), 1001-1013. https://doi.org/10.1098/rstb.2005.1634

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, *34*(4), 537-541.

Bouveyron, C., Celeux, G., & Girard, S. (2011). Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, *32*(14), 1706-1713. doi:https://doi.org/10.1016/j.patrec.2011.07.017.

Calhoun, V. D., Liu, J., & Adalı, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*, *45*(1), S163-S172.

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, *33*(2), 201-208. https://doi.org/10.1016/S0167-7152(96)00128-9

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.

De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., & Onghena, P. (2011). Clusterwise Simultaneous Component Analysis for Analyzing Structural Differences in Multivariate Multiblock Data. *Psychological Methods*. doi:10.1037/a0025385

De Vos, F., Koini, M., Schouten, T. M., Seiler, S., van der Grond, J., Lechner, A., ... & Rombouts, S. A. R. B. (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *Neuroimage*, *167*, 62-72. doi:10.1016/j.neuroimage.2017.11.025

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., ... & Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, *23*(1), 28-38.

Dubbelink, K. T. O., Stoffers, D., Deijen, J. B., Twisk, J. W., Stam, C. J., Hillebrand, A., & Berendse, H. W. (2013). Resting-state functional connectivity as a marker of disease progression in Parkinson's disease: a longitudinal MEG study. *NeuroImage: Clinical*, *2*, 612-619.

Durieux, J. (2021). *Clusterwise Independent Component Analysis R package*. Github. https://github.com/jeffreydurieux/CICA

Durieux, J., Rombouts, S. A. R. B., de Vos, F., Koini, M., & Wilderjans, T. F. (2021). Clusterwise Independent Component Analysis (C-ICA) for multi-subject resting-state fMRI data: Clustering subjects to capture heterogeneity in underlying functional connectivity patterns. *Manuscript submitted for publication.*

Fox, M. D., & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, *8*(9), 700-711.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, *6*(1), 99-106. doi:https://doi.org/10.2307/3236261

Gili, T., Cercignani, M., Serra, L., Perri, R., Giove, F., Maraviglia, B., ... & Bozzali, M. (2011). Regional brain atrophy and functional disconnection across Alzheimer's disease evolution. *Journal of Neurology, Neurosurgery & Psychiatry*, *82*(1), 58-66. doi:10.1136/jnnp.2009.199935

Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., ... & Schatzberg, A. F. (2007). Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological Psychiatry*, *62*(5), 429-437. doi:10.1016/j.biopsych.2006.09.020

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning.* Springer, New York doi:10.1007/b94608

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112, p. 18). Springer, New York.

Jolliffe, I. (2005). Principal Component Analysis. In *Encyclopedia of Statistics in Behavioral Science.* Springer, New York.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20 (*1), 141-151.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.

Li, Y. O., Adalı, T., & Calhoun, V. D. (2007). Estimating the number of independent components for functional magnetic resonance imaging data. *Human Brain Mapping*, *28*(11), 1251-1266. doi:10.1002/hbm.20359

Luteijn, F., & Barelds, D. (2013). *Psychologische Diagnostiek in de Gezondheidszorg.* Den Haag: Boom Lemma uitgevers.

Lynall, M. E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., & Bullmore, E. (2010). Functional connectivity and brain networks in schizophrenia. *Journal of Neuroscience*, *30 (*28), 9477-9487. doi:10.1523/JNEUROSCI.0333-10.2010

Meachler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). Cluster: Cluster Analysis Basics and Extensions. R Package version 2.1.2.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190-204.

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199-203. doi:10.1002/wics.199

R Core Team. (n.d.). A language environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/.

Rossbroich, J., Durieux, J., & Wilderjans, T. F. (2022). Model Selection Strategies for Determining the Optimal Number of Overlapping Clusters in Additive Overlapping Partitional Clustering. *Journal of Classification*, 1-38.

Richardson, M. (2009). Principal component analysis. Retrieved from http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA.pdf

Särelä, J., & Vigário, R. (2003). Overlearning in marginal distribution-based ICA: analysis and solutions. *The Journal of Machine Learning Research*, *4*, 1447-1469.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664-681.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

Timmerman, M. E., & Kiers, H. A. (2000). Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima. *British Journal of Mathematical and Statistical Psychology*, *53*(1), 1-16.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611-622.

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). CHull: A generic convex-hull-based model selection method. *Behavior Research Methods*, *45*(1), 1-15. doi: 10.3758/s13428-012-0238-5

Zhang, J., Kucyi, A., Raya, J., Nielsen, A. N., Nomi, J. S., Damoiseaux, J. S., ... & Whitfield-Gabrieli, S. (2021). What have we really learned from functional connectivity in clinical populations? *NeuroImage*, *242*, 118466.

Appendix A

This appendix consists of seven different tables. The first five tables show the mean estimation errors for the simultaneous selection methods using the five different complexity measures (see Table 2 in the main text). Table A1-A5 display the results for complexity measure C1-C5 respectively. Below an overview is given with the computations of the several complexity measures:

- C1: $Q + R$
- C2: $Q \times R$
- C3: $(T \times Q) + (I_r \times R)$
- C4: $(T \times Q) + (I_r \times R) + (Q \times R)$
- C5: $\left(T \times \frac{V}{T} \times Q\right) + (I_r \times R)$

Q equals the number of components, R number of clusters, T number of timepoints, I_r the total number of subjects per cluster in the data and V the number of voxels.

Table A6 shows the percentage of totally correct estimated models (i.e., correct estimate of the number of components AND the number of clusters) by the different simultaneous selection methods, presented for each complexity measure separately. Lastly, Table A7 shows the percentage of totally correct estimated models for the sequential methods.

Table A1

Mean estimation error (with SD) for the simultaneous model order selection methods ***using complexity measure 1****, computed overall and per level of the manipulated factors*

	Level	CHull	AIC	CHull	AICc	CHull	BIC	CHull	KIC	CHull	MDL	CHull	Overall
				AIC		AICc		BIC		KIC		MDL	
Number of components	2	0.11*	0.79	0.11*	0.79	0.11*	0.79	0.11*	0.79	0.11*	0.79	0.11*	0.42
		(0.11)	(0.10)	(0.11)	(0.10)	(0.11)	(0.10)	(0.11)	(0.1)	(0.11)	(0.1)	(0.11)	(0.36)
	5	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.35
		(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.39)
	25	0.00*	0.63	0.00*	0.63	0.00*	0.63	0.00*	0.63	0.00*	0.63	0.00*	0.29
		(0.00)	(0.11)	(0.00)	(0.11)	(0.00)	(0.11)	(0.00)	(0.11)	(0.00)	(0.11)	(0.00)	(0.32)
	50	0.15*	0.46	0.19	0.46	0.19	0.46	0.19	0.46	0.19	0.46	0.19	0.31
		(0.19)	(0.10)	(0.20)	(0.10)	(0.20)	(0.10)	(0.20)	(0.10)	(0.20)	(0.10)	(0.20)	(0.21)
Number of clusters	2	0.07*	0.76	0.09	0.76	0.09	0.76	0.09	0.76	0.09	0.76	0.09	0.40
		(0.15)	(0.14)	(0.17)	(0.14)	(0.17)	(0.14)	(0.17)	(0.14)	(0.17)	(0.14)	(0.17)	(0.37)
	4	0.06*	0.57	0.06*	0.57	0.06*	0.57	0.06*	0.57	0.06*	0.57	0.06*	0.29
		(0.10)	(0.13)	(0.09)	(0.13)	(0.09)	(0.13)	(0.09)	(0.13)	(0.09)	(0.13)	(0.09)	(0.28)
Amount of noise	.1	0.03*	0.67	0.06	0.67	0.06	0.67	0.06	0.67	0.06	0.67	0.06	0.33
		(0.07)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.34)
	.4	0.08*	0.67	0.08*	0.67	0.08*	0.67	0.08*	0.67	0.08*	0.67	0.08*	0.35
		(0.14)	(0.16)	(0.14)	(0.16)	(0.14)	(0.16)	(0.14)	(0.16)	(0.14)	(0.16)	(0.14)	(0.33)
	.7	0.09*	0.66	0.09*	0.66	0.09*	0.66	0.09*	0.66	0.09*	0.66	0.09*	0.35
		(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.33)
Overall		**0.06**	0.66	0.07	0.66	0.07	0.66	0.07	0.66	0.07	0.66	0.07	0.34
		(0.13)	(0.17)	(0.14)	(0.17)	(0.14)	(0.17)	(0.14)	(0.17)	(0.14)	(0.17)	(0.14)	(0.33)

Note. The method with the lowest overall mean estimation error is indicated in bold; the best method(s) per data characteristic is/are indicated with a *.

Table A2

Mean estimation error (with SD) for the simultaneous model order selection methods using ***complexity measure 2****, computed overall and per level of the manipulated factors*

	Level	CHull	AIC	CHull	AICc	CHull	BIC	CHull	KIC	CHull	MDL	CHull	Overall
				AIC		AICc		BIC		KIC		MDL	
Number of components	2	0.16*	0.79	0.17	0.79	0.17	0.79	0.17	0.79	0.17	0.79	0.17	0.45
		(0.05)	(0.10)	(0.06)	(0.10)	(0.06)	(0.10)	(0.06)	(0.10)	(0.06)	(0.10)	(0.06)	(0.32)
	5	0.27	0.77	0.24*	0.77	0.24*	0.77	0.24*	0.77	0.24*	0.77	0.24*	0.48
		(0.13)	(0.10)	(0.11)	(0.10)	(0.11)	(0.10)	(0.11)	(0.10)	(0.11)	(0.10)	(0.11)	(0.29)
	25	0.32*	0.63	0.36	0.63	0.36	0.63	0.36	0.63	0.36	0.63	0.36	0.48
		(0.05)	(0.11)	(0.12)	(0.11)	(0.12)	(0.11)	(0.12)	(0.11)	(0.12)	(0.11)	(0.12)	(0.17)
	50	0.01*	0.46	0.18	0.46	0.18	0.46	0.18	0.46	0.18	0.46	0.18	0.29
		(0.05)	(0.10)	(0.20)	(0.10)	(0.20)	(0.10)	(0.20)	(0.10)	(0.20)	(0.10)	(0.20)	(0.22)
Number of clusters	2	0.13*	0.76	0.15	0.76	0.15	0.76	0.15	0.76	0.15	0.76	0.15	0.43
		(0.09)	(0.14)	(0.10)	(0.14)	(0.10)	(0.14)	(0.10)	(0.14)	(0.10)	(0.14)	(0.10)	(0.33)
	4	0.25*	0.56	0.32	0.56	0.32	0.56	0.32	0.56	0.32	0.56	0.32	0.42
		(0.16)	(0.13)	(0.14)	(0.13)	(0.14)	(0.13)	(0.14)	(0.13)	(0.14)	(0.13)	(0.14)	(0.19)
Amount of noise	.1	0.18	0.66	0.17*	0.66	0.17*	0.66	0.17*	0.66	0.17*	0.66	0.17*	0.40
		(0.15)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.29)
	.4	0.19*	0.67	0.26	0.67	0.26	0.67	0.26	0.67	0.26	0.67	0.26	0.44
		(0.14)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.27)
	.7	0.20*	0.66	0.27	0.66	0.27	0.66	0.27	0.66	0.27	0.66	0.27	0.44
		(0.14)	(0.17)	(0.12)	(0.17)	(0.12)	(0.17)	(0.12)	(0.17)	(0.12)	(0.17)	(0.12)	(0.25)
Overall		**0.19**	0.66	0.23	0.66	0.23	0.66	0.23	0.66	0.23	0.66	0.23	0.42
		(0.14)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.27)

Note. The method with the lowest overall mean estimation error is indicated in bold; the best method(s) per data characteristic is/are indicated with a *.

Table A3

Mean estimation error (with SD) for the simultaneous model order selection methods using ***complexity measure 3****, computed overall and per level of the manipulated factors*

	Level	CHull	AIC	CHull	AICc	CHull	BIC	CHull	KIC	CHull	MDL	CHull	Overall
				AIC		AICc		BIC		KIC		MDL	
Number of components	2	0.00*	0.79	0.00*	0.79	0.00*	0.79	0.00*	0.79	0.00*	0.79	0.00*	0.36
		(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.40)
	5	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.35
		(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.39)
	25	0.18	0.63	0.10*	0.63	0.10*	0.63	0.10*	0.63	0.10*	0.63	0.10*	0.35
		(0.18)	(0.11)	(0.15)	(0.11)	(0.15)	(0.11)	(0.16)	(0.11)	(0.16)	(0.11)	(0.16)	(0.29)
	50	0.28*	0.46	0.28*	0.46	0.28*	0.46	0.28*	0.46	0.28*	0.46	0.28*	0.36
		(0.14)	(0.10)	(0.15)	(0.10)	(0.15)	(0.10)	(0.15)	(0.10)	(0.15)	(0.10)	(0.15)	(0.16)
Number of clusters	2	0.17	0.76	0.14*	0.76	0.14*	0.76	0.14*	0.76	0.14*	0.76	0.14*	0.42
		(0.20)	(0.14)	(0.19)	(0.14)	(0.19)	(0.14)	(0.19)	(0.14)	(0.19)	(0.14)	(0.19)	(0.35)
	4	0.06	0.57	0.05*	0.57	0.05*	0.57	0.05*	0.57	0.05*	0.57	0.05*	0.29
		(0.09)	(0.13)	(0.09)	(0.13)	(0.09)	(0.13)	(0.09)	(0.13)	(0.09)	(0.13)	(0.09)	(0.28)
Amount of noise	.1	0.08	0.67	0.05*	0.67	0.05*	0.67	0.05*	0.67	0.05*	0.67	0.05*	0.33
		(0.15)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.34)
	.4	0.11	0.67	0.08*	0.67	0.08*	0.67	0.08*	0.67	0.08*	0.67	0.08*	0.35
		(0.16)	(0.16)	(0.14)	(0.16)	(0.14)	(0.16)	(0.14)	(0.16)	(0.14)	(0.16)	(0.14)	(0.33)
	.7	0.15*	0.66	0.16	0.66	0.16	0.66	0.16	0.66	0.16	0.66	0.16	0.39
		(0.17)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.30)
Overall		0.12	0.66	**0.09**	0.66	**0.09**	0.66	**0.09**	0.66	**0.09**	0.66	**0.09**	0.36
		(0.16)	(0.17)	(0.16)	(0.17)	(0.16)	(0.17)	(0.16)	(0.17)	(0.16)	(0.17)	(0.16)	(0.33)

Note. The method with the lowest overall mean estimation error is indicated in bold; the best method(s) per data characteristic is/are indicated with a *.

Table A4

Mean estimation error (with SD) for the simultaneous model order selection methods using ***complexity measure 4****, computed overall and per level of the manipulated factors*

	Level	CHull	AIC	CHull AIC	AICc	CHull AICc	BIC	CHull BIC	KIC	CHull KIC	MDL	CHull MDL	Overall
Number of components	2	0.00*	0.79	0.00*	0.79	0.00*	0.79	0.00*	0.79	0.00*	0.79	0.00*	0.36
		(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.40)
	5	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.77	0.00*	0.35
		(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.39)
	25	0.12	0.63	0.08*	0.63	0.08*	0.63	0.08*	0.63	0.08*	0.63	0.08*	0.33
		(0.18)	(0.11)	(0.15)	(0.11)	(0.15)	(0.11)	(0.15)	(0.11)	(0.15)	(0.11)	(0.15)	(0.30)
	50	0.18*	0.46	0.24	0.46	0.24	0.46	0.24	0.46	0.24	0.46	0.24	0.33
		(0.18)	(0.10)	(0.19)	(0.10)	(0.19)	(0.10)	(0.19)	(0.10)	(0.19)	(0.10)	(0.19)	(0.19)
Number of clusters	2	0.13*	0.76	0.13*	0.76	0.13*	0.76	0.13*	0.76	0.13*	0.76	0.13*	0.42
		(0.19)	(0.14)	(0.19)	(0.14)	(0.19)	(0.14)	(0.19)	(0.14)	(0.19)	(0.14)	(0.19)	(0.36)
	4	0.02*	0.57	0.03	0.57	0.03	0.57	0.03	0.57	0.03	0.57	0.03	0.27
		(0.06)	(0.13)	(0.07)	(0.13)	(0.07)	(0.13)	(0.07)	(0.13)	(0.07)	(0.13)	(0.07)	(0.29)
Amount of noise	.1	0.00*	0.67	0.05	0.67	0.05	0.67	0.05	0.67	0.05	0.67	0.05	0.32
		(0.00)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.17)	(0.13)	(0.34)
	.4	0.10	0.67	0.05*	0.67	0.05*	0.67	0.05*	0.67	0.05*	0.67	0.05*	0.34
		(0.16)	(0.16)	(0.13)	(0.16)	(0.13)	(0.16)	(0.13)	(0.16)	(0.13)	(0.16)	(0.13)	(0.34)
	.7	0.13*	0.66	0.14	0.66	0.14	0.66	0.14	0.66	0.14	0.66	0.14	0.38
		(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.31)
Overall		**0.07**	0.66	0.08	0.66	0.08	0.66	0.08	0.66	0.08	0.66	0.08	0.35
		(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.17)	(0.15)	(0.33)

Note. The method with the lowest overall mean estimation error is indicated in bold; the best method(s) per data characteristic is/are indicated with a *.

Table A5

Mean estimation error (with SD) for the simultaneous model order selection methods using ***complexity measure 5****, computed overall and per level of the manipulated factors*

	Level	CHull	AIC	CHull	AICc	CHull	BIC	CHull	KIC	CHull	MDL	CHull	Overall
				AIC		AICc		BIC		KIC		MDL	
Number of components	2	0.00*	0.79	0.00*	0.79	0.00*	0.55	0.01	0.79	0.00*	0.55	0.01	0.32
		(0.00)	(0.10)	(0.00)	(0.10)	(0.00)	(0.15)	(0.00)	(0.10)	(0.00)	(0.15)	(0.00)	(0.36)
	5	0.00*	0.77	0.00*	0.77	0.00*	0.54	0.03	0.77	0.00*	0.54	0.03	0.32
		(0.01)	(0.10)	(0.01)	(0.10)	(0.01)	(0.14)	(0.03)	(0.10)	(0.01)	(0.14)	(0.03)	(0.35)
	25	0.26*	0.63	0.32	0.63	0.32	0.40	0.34	0.63	0.32	0.40	0.34	0.42
		(0.10)	(0.11)	(0.09)	(0.11)	(0.09)	(0.05)	(0.12)	(0.11)	(0.09)	(0.05)	(0.12)	(0.17)
	50	0.29*	0.45	0.40	0.45	0.40	0.38	0.44	0.45	0.40	0.38	0.44	0.41
		(0.10)	(0.09)	(0.18)	(0.09)	(0.18)	(0.03)	(0.20)	(0.09)	(0.18)	(0.03)	(0.2)	(0.15)
Number of clusters	2	0.18*	0.76	0.23	0.76	0.23	0.40	0.27	0.76	0.23	0.40	0.27	0.41
		(0.19)	(0.15)	(0.24)	(0.15)	(0.24)	(0.00)	(0.27)	(0.15)	(0.24)	(0.00)	(0.27)	(0.30)
	4	0.10*	0.56	0.13	0.56	0.13	0.53	0.14	0.56	0.13	0.53	0.14	0.32
		(0.10)	(0.13)	(0.15)	(0.13)	(0.15)	(0.16)	(0.14)	(0.13)	(0.15)	(0.16)	(0.14)	(0.25)
Amount of noise	.1	0.14*	0.67	0.15	0.67	0.15	0.47	0.16	0.67	0.15	0.47	0.16	0.35
		(0.15)	(0.17)	(0.17)	(0.17)	(0.17)	(0.12)	(0.16)	(0.17)	(0.17)	(0.12)	(0.16)	(0.28)
	.4	0.12*	0.67	0.17	0.67	0.17	0.47	0.20	0.67	0.17	0.47	0.20	0.36
		(0.14)	(0.16)	(0.21)	(0.16)	(0.21)	(0.13)	(0.23)	(0.16)	(0.21)	(0.13)	(0.23)	(0.29)
	.7	0.16*	0.65	0.22	0.65	0.22	0.45	0.26	0.65	0.23	0.45	0.26	0.38
		(0.17)	(0.18)	(0.24)	(0.18)	(0.24)	(0.14)	(0.26)	(0.18)	(0.24)	(0.14)	(0.26)	(0.28)
Overall		**0.14**	0.66	0.18	0.66	0.18	0.46	0.20	0.66	0.18	0.46	0.20	0.36
		(0.15)	(0.17)	(0.21)	(0.17)	(0.21)	(0.13)	(0.22)	(0.17)	(0.21)	(0.13)	(0.22)	(0.28)

Note. The method with the lowest overall mean estimation error is indicated in bold; the best method(s) per data characteristic is/are indicated with a *.

Table A6

Percentage of totally correct selected models per (simultaneous) selection method, computed per complexity measure

Complexity measure	CHull	AIC	CHull AIC	AICc	CHull AICc	BIC	CHull BIC	KIC	CHull KIC	MDL	CHull MDL	Overall
1	77.50	0.00	72.08	0.00	72.08	0.00	72.08	0.00	72.08	0.00	72.08	39.81
2	23.83	0.00	14.04	0.00	14.04	0.00	14.04	0.00	14.04	0.00	14.04	8.55
3	63.75	0.00	70.83	0.00	70.83	0.00	70.83	0.00	70.83	0.00	70.83	37.99
4	**78.33**	0.00	76.25	0.00	76.25	0.00	75.83	0.00	75.83	0.00	75.83	41.67
5	45.83	0.00	45.83	0.00	45.83	0.00	2.92	0.00	45.83	0.00	2.92	17.20

Note. The best performing (simultaneous) method with its best complexity measure is indicated in bold.

Table A7

Percentage of totally correct selected models per (sequential) selection method, computed overall and per level of the manipulated factors

	Level	VAF scree Components-clusters	VAF scree Clusters-components	PCA-Gap	PCA-K means	pPCA-Gap	pPCA-K means
Number of components	2	98.33*	96.67	0.00	0.00	0.00	16.67
	5	50.00	96.67*	0.00	0.00	0.00	33.33
	25	91.67	96.67*	0.00	0.00	0.00	20.00
	50	93.33	100.00*	0.00	0.00	0.00	20.00
Number of clusters	2	95.83	98.33*	0.00	0.00	0.00	41.67
	4	70.83	96.67*	0.00	0.00	0.00	0.00
Amount of noise	.1	87.50	100.00*	0.00	0.00	0.00	50.00
	.4	87.50	98.75*	0.00	0.00	0.00	25.00
	.7	75.00	93.75*	0.00	0.00	0.00	0.00
Overall		83.33	**97.50**	0.00	0.00	0.00	22.73

Note. The method with the largest percentage correct estimated models is indicated in bold; the best method(s) per data characteristic is/are indicated with a *.