LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Stella Bollmann, Andreas Hölzl, Moritz Heene, Helmut Küchenhoff, Markus Bühner

Evaluation of a new k-means approach for exploratory clustering of items

Technical Report Number 182, 2015
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Abstract

For the exploratory analysis of survey data commonly the exploratory factor analysis (EFA) is used. However, EFA is known to exhibit some problems. The major mathematical issue is the factor indeterminacy. Further problems are for example its weak performance in small sample sizes ($n \leq 150$) and with high cross-loadings (e.g. Guadagnoli & Velicer, 1988; Sass, 2010; Wayne F. Velicer & Fava, 1998) as well as the general issue of the underlying measurement model including uncorrelated residual variances, what may be difficult to justify (Cudeck & Henly, 1991; MacCallum & Tucker, 1991; R. C. Tryon, 1959). The authors suggest two new k-means approaches as an alternative: *k-means scaled distance measure* (*sdm*) where items are represented in a coordinate system in a way so that their distance is based on one minus their correlation; and *k-means cor* where item inter-correlations are directly taken as the coordinate points of the items. These approaches were tested in a resampling with two real data sets and a traditional Monte Carlo simulation, as well as in a cross validation using confirmatory factor analysis (CFA). For *dimensionality assessment* the cluster validity coefficient Silhouette was used. In either analysis these approaches were compared to existing cluster analysis approaches and EFA. The authors conclude that the main advantage of the new approaches are (a) that cluster scores are determinate and (b) for *item assignment k-means sdm* obtains better results than EFA and other cluster analysis approaches. The authors therefor suggest to use a combination of EFA methods for *dimensionality assessment* and k-means for *item assignment*.

1. Introduction

When talking about exploratory structure detection in psychometric data, we usually refer to detection of homogenous groups of items of questionnaires, when no prior information about the test structure is present. Homogenous typically means that items within one group are highly interrelated. The most frequently used method for this process is the exploratory factor analysis (EFA). It is based on the idea that latent variables cause relations between test items, and that fewer than p underlying factors thus, can explain these relationships between p items. Hence, the aim of EFA is the detection of the 'true' factor model,

$$\boldsymbol{R = \Lambda \Phi \Lambda' + U^2}$$

whereby $\mathbf{R}$ is the pxp correlation matrix of the p observed variables, $\mathbf{U^2}$ is the diagonal pxp matrix of the unique variances, $\mathbf{\Lambda}$ is the pxk matrix of the factor loadings on the k < p factors, and $\mathbf{\Phi}$ stands for the kxk matrix of the factor inter-correlations.

The exploratory structure detection with EFA involves two steps: First the *dimensionality assessment*, i.e. the assessment of the number of factors m and second the *item assignment* to factors. The second step, the *item assignment*, is directly based on the estimation of the loading matrix obtained from the principal axis factoring (PAF) given the number of factors m. In numerous studies with simulated and real data it has proven to work sufficiently accurate (Arrindell & Van der Ende, 1985; Guadagnoli & Velicer, 1988; Mundfrom, Shaw, & Ke, 2005; Sass, 2010, 2010; Schmitt, 2011; Wayne F. Velicer & Fava, 1998). The *dimensionality assessment* has to be done by a separate preceded method. For example, Parallel Analysis (PA, Horn, 1965a) and Minimum Average Partial rule (MAP rule, Velicer, 1976) have been suggested for that purpose after their accuracy had been shown in several studies (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Patil, Singh, Mishra, & Todd Donavan, 2008; W. F. Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986).

However, EFA has been criticized for some unsolved problems it reveals. One huge mathematical drawback of EFA is the problem of factor indeterminacy. Even when the factor model is defined on a correlation level, it is not defined on the data level (e.g. Anderson, 1958). For the calculation of factor scores there are always infinite solutions that satisfy the equation system. Although the estimation of factor scores is possible (Schonemann & Steiger, 1978), the solution of the equation system is not determined.

Furthermore, some rather practical issues have been found. For example, EFA cannot cope with small samples of $n \leq 150$ (Guadagnoli & Velicer, 1988; Sass, 2010; Wayne F. Velicer & Fava, 1998) although they are not uncommon in psychological research (Fabrigar et al., 1999). In fact, Fabrigar et al. (1999) showed in their summary that 30% of the current studies in the use of FA, had used sample sizes of n=100 or less. Moreover, the presence of medium or high cross-loadings may cause several problems as for example biased estimates of parameters of the rest of the model depending on the chosen rotation criterion (Asparouhov & Muthén, 2009; Sass & Schmitt, 2010). Other objections are somehow more of a theoretical nature. For example, it had been noted that the assumption of observed values that are a linear composition of a systematic part that is explained by a few underlying factors plus an error term that has to be uncorrelated is very simplified and fails to meet the criterion of real live circumstances (R. C. Tryon, 1935, 1959). Cattell (1987, S101 ff.) was already aware that the data are in fact the result of many more underlying, systematic factors than observed variables. Until today researchers keep pointing out that no factor model is entirely true, even in the population (Cudeck & Henly, 1991; MacCallum & Tucker, 1991). While most researchers nevertheless continue to use EFA in order to find the model that best fits the data, the question arises whether there is not a different way. It would be desirable to reduce dimensionality and to assign items to subtests without the need of a complex underlying model about the composition of item scores. For this purpose, Cluster Analysis (CA) has been suggested for the use of

clustering of items more than 70 years ago (R. C. Tryon, 1939). Since then, a controversial debate has been going on whether CA could be an alternative to EFA or not. A quite popular quote in that context is that from Tryon and Bailey (1970): "Cluster analysis is a poor man's factor analysis". In contrast, a number of researchers have kept suggesting complementing the process of assigning items to sub tests with CA or even substituting EFA with CA. They proposed several different hierarchical and non-hierarchical CA methods (Bacon, 2001; Hunter, 1973; Loevinger, Gleser, & DuBois, 1953; W. Revelle, 1979; Schweizer, 1991). Hunter (1973) for example, noticed that EFA is not a useful instrument for assessing the homogeneity of an item pool and therefore suggested to add a subsequent CA to every EFA. He therefore proposed the similarity coefficient. Revelle (1979) presented the ICLUST method using the reliability coefficient beta for determination of the number of clusters. It conservatively estimates the homogeneity by using the lowest possible split-half reliability of the scale. Revelle (1979) could show that results he obtained were more useful than those from EFA. Schweizer (1991) then used the hierarchical CA with disaggregated correlations and Bacon (2001) developed the correlations of correlations as a distance measure. Loevinger et al. (1953) were the first ones to use a non-hierarchical approach in psychometrics. There, triplets of items are detected that have the highest similarity according to a previously chosen similarity matrix. Then, iteratively items are added to the cluster that increases the homogeneity of the cluster and those items are removed that decrease the homogeneity of the cluster. This procedure is then repeated for the remaining items until all items are assigned to a cluster. None of these researchers though, has tried to use k-means for clustering of items. The present study aims to find an algorithm that clusters items by applying the k-means approach directly. In Section 2, we present possible methods for non-hierarchical clustering of items and in Section 3, we compare these methods with hierarchical cluster methods using real data and simulated data.

2. Non-hierarchical clustering of items

We will first explain an existing approach for the clustering of variables that so far has not been used in a psychometric context. This approach assumes the centre of the cluster to be a latent variable. Then, two new k-means approaches will be introduced that we developed for clustering of items.

2.1. *ClustOfVar*

ClustOfVar (Chavent, Kuentz, Liquet, & Saracco, 2011) is a clustering technique that has mainly been developed for the purpose of dimension reduction and variable selection for mixtures of quantitative and qualitative variables (Chavent et al., 2011) and has been applied on gene expressions data (Chavent, Genuer, Kuentz-Simonet, Liquet, & Saracco, 2013). The idea is to define a synthetic variable as the mean of the cluster. In the case of metric variables the synthetic variable is that variable of which the sum of squared correlations to all variables of the cluster is maximal. It can be shown (Vigneau & Qannari, 2003) that this synthetic variable is the first principal component of all the variables in the cluster. The sum of squared correlations of all variables to the synthetic variable is referred to as homogeneity (H). Based on these two definitions two cluster algorithms are defined:

1. A hierarchical cluster method and
2. A k-means cluster method.

The first one is an agglomerative hierarchical cluster method with the distance measure:

$$d(A,B) = H(A) + H(B) - H(A \cup B)$$

where in each step those two clusters are merged that have the smallest d. The number of clusters can be determined e.g. via examination of the dendrogram.

In the k-means approach an initial clustering is selected which can be either at random or obtained from the hierarchical cluster approach. Based on this initial clustering a synthetic variable, which is the first principal component of the correlation matrix of all variables of the respective cluster, is calculated for each cluster. Then, the variables are assigned to the cluster

with which mean they have the highest correlation. This process is optimized iteratively until convergence is reached.

2.2. K-means clustering of items

The idea of k-means clustering of items is that the items are points in a coordinate system and the square distances between each item of a cluster and its centre are minimized iteratively. It can be expressed as follows:

$$\min_{\{m_k\},1\le k\le K}\sum_{k=1}^{K}\sum_{x\in C_k}(\mathbf{x}-\mathbf{m}_k)^2$$

where $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ is the data to be clustered, $\mathbf{m}_k$ is the centroid of cluster C_k and K is the previously defined number of clusters. As a baseline condition k random items are each assigned to one cluster, so that each cluster consists of exactly one item. Then the item assignment is optimized iteratively until the result no longer changes significantly.

Beforehand, a way has to be found in which the items can be represented in a coordinate system. For clustering of items naturally the principal aim is to cluster those items together that have the highest correlation. For this reason, both methods implemented here are based on correlations.

2.2.1. k-means scaled distance measure (sdm).

In the first approach, the idea is to create coordinate points in a way so that the distance between these points are directly based on correlations, or more precisely on 1 minus the correlation (*1-cor*). As a result those items are close to each other that are highly correlated. However, *1-cor* is not a metric (van Dongen & Enright, 2012) and therefore this distance measure cannot be interpreted as a Euclidian distance between two points in a space. Van Dongen and Enright (2012) though could show that

$$d(A,B)=\sqrt{0.5-0.5\cdot Cor(A,B)}$$

is a metric and therefore this distance measure shall be used in this paper. To represent the items in a coordinate system in a way so that the distance between two items is d(A,B), a

scaling procedure is used that is based on the idea of multidimensional scaling. Given that we have *n* items, the solution shall have *n-1* dimensions. Multidimensional scaling is a method that estimates a matrix of coordinate points with a preliminary specified number of dimensions (D^2) from a distance matrix (X). X is chosen such that it is a column centered matrix with the column sum scores being *0*. With the distances of each of the n points to each other given, one has $n \cdot (n-1)$ equations for the distances. Given that each point has a dimension of (n-1) we obtain $n \cdot (n-1)$ unknown coordinates and thus the solution is unique. How the scaling of the matrix is done in detail can be found in Borg & Groenen (2005).

2.2.2. k-means correlation (cor).

The idea of the second approach is that the correlations between the items are directly used as the coordinates of the items. The vector of correlations of one item to all the other items is represented as the coordinates of this item. The distance between two items is then

$$d(A,B) = \sum_i (Cor(A,V_i) - Cor(B,V_i))^2$$

All k-means approaches have one advantage over hierarchical methods: They allow us to cluster items around a centroid. Vigneau and Qannari (2003) called this centroid the "synthetic variable" what can also be regarded as the construct to be measured. As a consequence we obtain distances between items and the centroid of their cluster, e.g. the construct - what is equivalent to a loading matrix in EFA. However, there is no need to establish a whole measurement model that includes assumptions about uncorrelated residual variances.

Just like in EFA in CA the *item assignment* to clusters has to be preceded by a *dimensionality assessment*. When using hierarchical clustering, this decision usually is based on information derived from the dendrogram. This is not possible when using non-hierarchical procedures. Here, usually the user has to specify the number of clusters before

any clustering is accomplished (Milligan & Cooper, 1985). In other disciplines that use non-hierarchical clustering, like genomic research, *dimensionality assessment* is commonly made using visual inspection and prior knowledge while the fit with the data is less frequently examined (Handl, Knowles, & Kell, 2005). Handl et al. (Handl et al., 2005) give an overview of cluster validation techniques that can be used for determination of the best-fitting number of clusters and some of them are implemented in the R package "clValid" (Brock, Pihur, Datta, & Datta, 2008). One group of validation methods they use was the internal measures that measure internal validity in a way so that they "reflect the compactness, connectedness, and separation of the cluster partitions" (Brock et al., 2008). One of these methods is also used in the following.

2.3 Silhouette width for assessment of dimensionality

The Silhouette width is based on each observation's Silhouette value, which is defined in the following manner:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}$$

with a_i being the average distance between *i* and all other observations in the same cluster, and b_i is the average distance between *i* and the observations in the "nearest neighboring cluster". A high Silhouette value means a good fit of the item to the respective cluster and a low fit to all the other clusters.

In the case of clustering of items, one silhouette value is obtained for each item and these values are then all summed up in order to get the Silhouette width. When a Silhouette width is computed for the suggested cluster solution of each possible number of clusters it is possible to select that number of clusters for which the Silhouette width is highest. This method will be referred to as Silhouette.

3. Comparison of methods

The purpose of this study was first to examine new k-means approaches that have not been used for clustering of variables before. These approaches are *k-means sdm* and *k-means cor*. Furthermore, two other variable clustering approaches that have only been used for genomic data shall be applied to psychometric data, *ClustOfVar I*, which is a hierarchical approach and *ClustOfVar II*, the k-means approach. Their comparative performance to EFA and traditional hierarchical CA methods shall be investigated. The two hierarchical clustering methods used, is one with complete linkage (CACL) and one with average linkage (CAAL), both with the distance measure:

$$d(A,B) = 1 - cor(A,B)$$

For EFA, one PAF is conducted with pearson correlations, promax rotation and ML estimation. Van der Linden et al. (2012) showed in a study on factor inter-correlations of different personality inventories that factors are typically correlated. Their inter-correlations range from .52 to .67 (in absolute values) and the average inter-correlation is .60. For these reasons, we chose oblique Promax rotation for our study. Items are assigned to those factors on which they have the highest absolute loading.

Second, a formula for computation of cluster validity, Silhouette, shall be used as a stopping rule and its performance will be investigated. Its performance in *dimensionality assessment* is compared to the performance of three established methods of *dimensionality assessment* in EFA: parallel analysis with principal component analysis (PA-PCA) and with principal axis factoring (PA-PAF) according to Horn (1965) and the minimum average partial rule (MAP rule) by Velicer (1976). They have been chosen because their accuracy compared to other methods has been shown in previous studies and they can easily be automated for use in simulation studies (Fabrigar et al., 1999; Patil et al., 2008; W. F. Velicer et al., 2000; Zwick & Velicer, 1986). The MAP was based on a PAF with maximum likelihood estimation and Promax rotation. For all CA methods the number of clusters was determined with Silhouette.

3.1 Design of the simulation study

In order to obtain a comprehensive image of the accuracy of these methods three different types of simulations will be used. First, we will use the Real World simulation, a new approach using real data. Second, a Traditional simulation study is conducted where a few variables are varied and different sample sizes are drawn. And third, we will make a cross validation with confirmatory factor analysis (CFA).

3.1.1. Real World simulation

For comparison of the performance of the methods, first a resampling technique is used that is referred to as Real World simulation. The result of each method in a real, huge data set is regarded as the population model. Samples of sizes 100, 200, 500, 1000 are then drawn with replacement from this data set and the results of each method in the sub samples are compared to the population model. Each sample size is replicated 1000 times. Thus, we have 4 x 4 different conditions for determining the number of factors as we compare 4 different methods (Silhouette, PA-PAF, PA-PCA and MAP rule) and 7 x 4 for *item assignment*. The 7 *item assignment* methods we compare are: k-means scaled distances, *k-means cor*, *ClustofVar* hierarchical, *ClustofVar* k-means, CACL, CAAL and EFA.

1. NEO-PI-R data. NEO-PI-R is a widely used personality inventory measuring personality in five major domains (Ostendorf & Angleitner, 2004): neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Each domain scale is divided into six facets and eight items operationalize each facet. Thus, the questionnaire consists of 240 items. For this study, a self-report form was used in which participants provided self-reports on typical behaviors or reactions on a five point Likert scale, ranging from 0='strongly disagree' to 4='strongly agree'. Validity and reliability for all domain scales was shown by Ostendorf and Angleitner (2004). For the present paper, we took all the items of one sub-facet per factor. We assigned the resulting 40 items to the overlying 5 sub-facets. We chose the

following 5 facets from each of the 5 factors: N1 (Anxietey), E2 (Gregariousness), O3 (Feelings), A4 (Compliance) and C5 (Self-discipline). As indicated in the manual of the NEO-PI-R the correlation matrix of items shows intermediate intercorrelations of items within facets (from .18 to .36) and low intercorrelations between factors (for more details see Ostendorf & Angleitner, 2004). Visualisations of both correlations matrices are provided in Figure 1. The mean of factor inter-correlations when specifying the theoretical five-factor model is .00 and the mean of absolute values of factor inter-correlations is .13 ranging from .01 to .33. Main factor loadings are between .28 and .79. (see Table 1). Although these values may seem particularly low, they are rather typical for personality questionnaires. Peterson (2000) showed in a meta-analysis on factor loadings in EFA's of questionnaire data, that the average factor loading is .32 with 25% of the factor loadings being less than .23, and 25% greater than .37. Cross-loadings in the NEO-PI-R data show a mean of .00, ranging from -.21 to .16. All in all the NEO-PI-R data set exhibits relatively low cross loadings and low factor inter-correlations (see Table 2). The NEO-PI-R norm data set consists of 11,724 participants. The mean age of the sample is 29.92, ranging from 16 to 91 with 36% males and 64% females.

2. IST-2000-R data. The basic module of the IST-2000-R measures intelligence in three major domains: verbal intelligence, numerical intelligence and spatial intelligence each of which is divided into three sub-tests (Amthauer, Brocke, Liepmann, & Beauducel, 2007). The test compromises a total of 180 questions, which can only be answered true or false. The sub-tests were the basis of our calculations. We treated them as variables of our data set and assembled them into overlying factors. This was done in order to avoid the problem of binary data that will be addressed in further research. Main factor loadings range from .47 to .83 (see Table 3). Validity and reliability for all sub-tests was shown by Amthauer et al. (2007). Cross-loadings range from -.11 to .18 (see Table 3) with a mean of .01. The three factor inter-correlations

Table 1
Loading Matrix of Population Data Set NEO-PI-R

	N1	E2	O3	A4	C5
V1	.62				
V31	.59				
V61	.75				
V91	.52				
V121	.53				
V151	.67				
V181	.68				
V211	.47		.13	-.11	
V7		.61			
V37		.79			
V67		.50		.14	
V97	.13	.51	.11		
V127		.43		.13	
V157		.43	-.21		
V187		.59			
V217		.75		-.15	
V13			.56	-.10	
V43			.62		
V73			.56		
V103			.39		
V133			.71		
V163			.58		
V193			.42		
V223			.46		
V19			.14	.28	
V49	.11		-.13	.41	.14
V79		-.18		.42	
V109				.46	
V139				.35	
V169		-.13		.54	
V199				.54	
V229	-.17			.38	.16
V25					.64
V55					.72
V85	.10				.63
V115					.69
V145					.57
V175	-.17		.11		.39
V205					.58
V235					.55

Note. N1=Anxietey; E2=Gregariousness; O3=Feelings; A4=Compliance; C5=Self-discipline. Promax rotation; ML estimation. Loadings below .10 are supressed.

have the values .66, .49, and .43. To sum it up, the level of cross-loadings is comparable to that in the NEO-PI-R data set and factor inter-correlations are much higher. The norm data set consists of 1,352 observations. The mean age is 19.09 ranging from 16 to 25 with 44% females and 56% males.

For *dimensionality assessment,* success rates were reported for each sample size, i.e. the percentage of identical number of factors as in the population data. And for *item assignment,* we set the number of factors to the theoretically assumed number: 5 factors for the NEO-PI-R and 3 factors for the IST-2000-R. Similarity was then determined using the

Table 2
Loading Matrix of Population Data Set and Population Model IST-2000-R

Facet	Factor 1	Factor 2	Factor 3
V1		.74	-.11
V2		.67	.18
V3	.18	.53	
N1	.47	.12	.18
N2	.83		
N3	.74		
F1			.64
F2			.49
F3	-.11		.51

Note. V=Verbal Intelligence; N=Numerical Intelligence; F=Figural Intelligence. Promax rotation; ML estimation. Loadings below .10 are suppressed.

Table 3
Factor Inter-correlation Matrix of the Population Data Set NEO-PI-R

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Factor 1	1.0	-.33	-.14	.19	-.06
Factor 2		1.0	.01	-.08	.07
Factor 3			1.0	.28	.01
Factor 4				1.0	-.03
Factor 5					1.0

Note. Promax roation; ML estimation.

Rand Index (Rand, 1971), calculated by counting the number of correctly classified pairs of elements. The Rand Index is defined by:

$$R(C, C') = \frac{2(n_{11} + n_{00})}{n(n-1)}$$

C is the actual cluster solution in the sample, C' is the cluster solution in the population data set, n_{11} is the number of pairs that are in the same cluster under C and C', and n_{00} is the number of pairs that are in different clusters under C and C'.

3.1.2. Traditional Simulation

We specified the factor model, the EFA had found in each of the population data sets. Estimates of main and cross loadings, factor intercorrelations and uniquenesses were obtained from the factor analysis of the norm data set and subsequently used for the simulations (population model). In the first simulation condition, all residual correlations were set to zero, which means that a perfectly fitting model was simulated. And in the second simulation condition, residual correlations also were obtained from the population data and included into the simulation.

Additionally, we also applied different sample sizes (100, 500 and 1000) with 1000 replications each. Success rates in the sub-samples are reported for *dimensionality assessment* and Rand Indexes are reported for *item assignment*. We therefore had 5 x 3 x 2 different conditions for determining the number of factors as we compared 5 different methods with 3 sample sizes and 2 simulation conditions and 7 x 3 x 2 for *item assignment*.

Sample covariance matrices were drawn from the given population covariance matrix (calculated from loadings, factor inter-correlations and residual variances). On basis of these sample covariance matrices, we applied the different methods.

3.1.3. Cross validation with CFA

We specified factor models with all combinations of each of the *dimensionality assessment* methods and the *item assignment* methods on sub-samples of the data set with n=100, 500, 1000 and 1000 replications each. We subsequently tested the specified model on the entire data set with a CFA. We then reported BIC values of each model.

All calculations were programmed in the open source software R 0.94.110.

3.2. Results

3.2.1. Real World simulation

Dimensionality assessment. Table 4 summarizes the results of the Real World simulation *dimensionality assessment* first for the NEO-PI-R data set and second for the IST-2000-R data

Table 4

Success Rates for dimensionality assessment in the Real World simulation for Different Sample Sizes for both data sets

Method		# pop data	n			
			100	200	500	1000
NEO-PI-R						
K-means sdm	Silhouette	6	.28	.35	.41	.60
K-means cor	Silhouette	5	.58	.83	.98	1.00
ClustOfVar I	Silhouette	5	.65	.84	.96	.99
ClustOfVar II	Silhouette	5	.69	.89	.96	1.00
CAAL	Silhouette	6	.23	.31	.49	.60
CACL	Silhouette	5	.14	.37	.66	.88
EFA	PA-PAF	8	.02	.00	.00	.01
	PA-PCA	6	.26	.34	.53	.89
	MAP	5	.56	.73	.90	.96
IST-2000-R						
K-means sdm	Silhouette	2	.65	.80	.88	.98
K-means cor	Silhouette	2	.71	.80	.93	.96
ClustOfVar I	Silhouette	2	.64	.76	.90	.96
ClustOfVar II	Silhouette	2	.67	.75	.87	.97
CAAL	Silhouette	2	.68	.72	.82	.89
CACL	Silhouette	2	.58	.68	.86	.93
EFA	PA-PAF	4	.18	.27	.39	.64
	PA-PCA	2	.58	.80	.95	.99
	MAP	1	.99	1.00	1.00	1.00

Note. K-means sdm=k-means scaled distance measure; k-means cor=k-means correlation; *ClustOfVar I=ClustOfVar* with hierarchical cluster analysis; *ClustOfVar II=ClustOfVar* with k-means cluster analysis; CAAL=hierarchical cluster analysis with average linkage; CACL=hierarchical cluster analysis with complete linkage; PA-PAF= Parallel analysis with Principal axis factoring, PA-PCA=Parallel analysis with Principal component analysis; MAP=MAP rule; n=sample size; # pop data= number of factors in population data set. All success rates are based on 1000 replications.

set. The column "# pop data" shows how many factors or clusters the respective method suggested in the population data set. The following columns display the success rates, i.e. the percentages of identical number of dimensions as indicated in the population data for different

sample sizes. The methods with the highest success rates are *k-means cor* with Silhouette in both data sets, *ClustOfVar II* in the NEO-PI-R data set and the EFA methods PA-PCA, MAP in the IST-2000-R data set. Still, the high success rate of MAP in the IST-2000-R data set is at the expense of a very uninformative factor solution of one common factor. The hierarchical CA methods (CAAL and CACL) had the lowest success rates across data sets. Overall, in the IST-2000-R data only the PA methods were sensitive for sample size (see Table 4) but not the CA methods.

Item assignment. Next, Table 5 show the mean proportions of *item assignments* identical to the *item assignment* of the respective method (Rand Index) in the population data for different sample sizes for both data sets separately. The new *k-means sdm* shows the highest Rand Index of all methods in both data sets. Especially for smaller sample sizes of 100 and 200 it outperforms EFA. Also *ClustOfVar II* achieved at least as high a Rand Index as EFA in both data sets. The lowest proportions of identical solutions were obtained by the traditional hierarchical methods.

Please note that *ClustOfVar I* and *ClustOfVar II* often yield the same results. This is because *ClustOfVar I* builds on the results of *ClustOfVar I* and it is possible that these first results cannot be further improved.

3.2.2. Traditional simulation

Dimensionality assessment. Table 6 summarizes the results of the Traditional simulation for *dimensionality assessment* for the two different simulation conditions. The success rates and means of indicated numbers of dimensions for results from the simulated samples in comparison to the simulated model are reported. When specifying the population model without residual correlations, EFA methods were the ones to achieve the highest success rates. When adding residual correlations though, their performance dropped considerably while other methods namely *k-means cor* and *ClustOfVar* perform better. In the IST-2000-R data set almost no method could find the specified three factors. Mostly, two

Table 5
Rand Indexes in the Real World simulation for different item assignment methods for different sample sizes for both data sets

n	K-means sdm	K-means cor	ClustOfVar I	ClustOfVar II	CAAL	CACL	EFA
			NEO-PI-R				
100	.96	.88	.94	.95	.86	.88	.94
200	.99	.89	.97	.97	.90	.93	.98
500	1.00	.89	.99	.98	.93	.97	1.00
1000	1.00	.89	.99	.99	.95	.98	1.00
			IST-2000-R				
100	.84	.78	.83	.83	.72	.77	.78
200	.87	.81	.86	.86	.73	.77	.83
500	.91	.83	.90	.90	.72	.74	.88
1000	.91	.85	.91	.91	.72	.73	.89

Note. K-means sdm=k-means scaled distance measure; k-means cor=k-means correlation; *ClustOfVar I=ClustOfVar* with hierarchical cluster analysis; *ClustOfVar II=ClustOfVar* with k-means cluster analysis; CAAL=hierarchical cluster analysis with average linkage; CACL=hierarchical cluster analysis with complete linkage; EFA=exploratory factor analysis. All Rand Indexes are based 1000 replications.

factors were suggested, what was the lowest possible number for the CA methods. These results are probably due to the high factor correlations in this data set that were included in the simulation. PA-PAF was the only method to find three factors fairly often.

Item assignment. In the NEO-PI-R data, all methods had slightly higher Rand Indexes in both traditional simulation conditions than in the Real World simulation (see Table 7). The highest Rand Indexes though were obtained by *k-means sdm* followed by EFA and *ClustOfVar*. However, EFA only was poorer than *k-means sdm* in the IST-2000-R data set. In the NEO-PI-R data set both *k-means sdm* and EFA achieved almost a 100% correct indications. The traditional hierarchical methods again showed the lowest Rand Indexes. Summing up the results from the Real World simulation and the Traditional simulation *k-means cor* with Silhouette PA-PCA performed best in assessing dimensionality across all conditions. On the other hand, for *item assignment* the new k-means approach *k-means sdm* obtained the best results especially for small sample sizes. According to these results, one

Table 6

Success Rates and Means of indicated Number of Factors across all Sample Sizes in the Traditional simulation for both Data Sets

Method		Population model		Population model + res cor	
		Success Rate	M	Success Rate	M
		NEO-PI-R			
K-means sdm	Silhouette	.48	7.65	.44	7.77
K-means cor	Silhouette	.89	3.83	.86	3.86
ClustOfVar I	Silhouette	.90	4.69	.89	4.83
ClustOfVar II	Silhouette	.92	4.57	.91	4.75
CAAL	Silhouette	.40	7.59	.37	7.64
CACL	Silhouette	.93	8.40	.98	8.52
EFA	PA-PAF	.99	5.36	.33	5.94
	PA-PCA	.98	5.05	.56	5.39
	MAP	.97	3.10	.85	3.08
		IST-2000-R			
K-means sdm	Silhouette	.05	2.66	.05	2.68
K-means cor	Silhouette	.05	2.31	.04	2.29
ClustOfVar I	Silhouette	.07	2.69	.06	2.68
ClustOfVar II	Silhouette	.07	2.70	.06	2.69
CAAL	Silhouette	.04	2.54	.09	2.59
CACL	Silhouette	.07	2.90	.06	2.95
EFA	PA-PAF	.92	3.17	.57	3.30
	PA-PCA	.04	1.65	.04	1.66
	MAP	.00	1.00	.00	1.00

Note. K-means sdm=k-means scaled distance measure; k-means cor=k-means correlation; *ClustOfVar I=ClustOfVar* with hierarchical cluster analysis; *ClustOfVar II=ClustOfVar* with k-means cluster analysis; CAAL=hierarchical cluster analysis with average linkage; CACL=hierarchical cluster analysis with complete linkage; PA-PAF= Parallel analysis with Principal axis factoring, PA-PCA=Parallel analysis with Principal component analysis; MAP=MAP rule; M=mean; res cor=residual correlations. Simulated number of factors: NEO-PI-R:5, IST-2000_r:3. All Success Rates are based on 1000 replications. Sample sizes= 100, 500, 1000.

Table 7
Rand Indexes across all sample sizes in the Traditional simulation for both data sets

	K-means sdm	K-means cor	ClustOfVar I	ClustOfVar II	CAAL	CACL	EFA
				NEO-PI-R			
Population model	.99	.97	.98	.99	.93	.96	.99
Population model + res cor	.99	.96	.97	.98	.92	.95	.99
				IST-2000-R			
Population model	.96	.91	.95	.95	.70	.77	.95
Population model + res cor	.96	.90	.94	.94	.70	.74	.93

Note. Sample size= 100, 500, 1000; number of replications= 1000.

could expect *k-means sdm* and PA-PCA to be the best combination for exploratory structure detection. The suggested models of all combinations of *dimensionality assessment* methods and *item assignment* methods were tested in a CFA cross validation.

3.2.3. CFA cross validation

The BIC's in the CFA cross validation in NEO-PI-R were all above 930,000 and in IST-2000-R above 62,000. We therefore report the obtained BIC's minus 930,000 and minus 62,000 respectively in Table 8. According to expectations, the combination of *k-means sdm* and PA-PCA obtained the lowest BIC in IST-2000-R followed by the non-hierarchical CA methods together with PA-PCA. In NEO-PI-R though, PA-PAF together with *ClustOfVar II* and *k-means sdm* showed the best results. In NEO-PI-R, also Silhouette with *k-means sdm* had a low BIC. Altogether, PA combined with either *ClustofVar II* or *k-means sdm* suggested the best fitting factor models according to CFA. PA-PCA turned out to be more useful in the IST-2000-R data set where high factor-intercorrelations are present whereas PA-PAF was more useful in the NEO-PI-R data set. In the IST-2000-R data, MAP always resulted in the same BIC, for the reason that it always indicated one factor. Based on the results previously mentioned, high chances for a good overall result could also be expected from a combination

of *k-means cor* with silhouette for *dimensionality assessment* and *k-means sdm* for *item assignment*, which was not tested in the CFA cross validation.

Table 8
BIC for CFA cross validation with different combinations of dimensionality assessment and item assignment methods for both data sets

	K-means sdm	K-means cor	ClustOfVar I	ClustOfVar II	CAAL	CACL	EFA
			NEO-PI-R				
Silhouette	5477	12580	7382	6614	6053	7060	-
PA-PAF	5406	6659	5870	5372	11384	10193	6200
PA-PCA	5828	7795	6494	5913	13708	11579	6636
MAP	6136	8370	6882	6287	14568	12078	6936
			IST-2000-R				
Silhouette	1857	1832	1854	1854	1911	1884	-
PA-PAF	1936	1941	1934	1934	2050	2077	1884
PA-PCA	1810	1816	1810	1811	1873	1838	1830
MAP	1964	1964	1964	1964	1964	1964	1964

Note. K-means sdm=k-means scaled distance measure; k-means cor=k-means correlation; *ClustOfVar I=ClustOfVar* with hierarchical cluster analysis; *ClustOfVar II=ClustOfVar* with k-means cluster analysis; CAAL=hierarchical cluster analysis with average linkage; CACL=hierarchical cluster analysis with complete linkage; PA-PAF= Parallel analysis with Principal axis factoring, PA-PCA=Parallel analysis with Principal component analysis; MAP=MAP rule. Number of replications= 1000. Reported values: BIC minus 930,000 for NEO-PI-R and BIC minus 62,000 for IST-2000-R. Sample size= 100, 500, 1000.

4 Discussion

The above results, using Real World simulation, Traditional Simulation and a CFA cross validation, show that the two new k-means approaches *k-means sdm* and *k-means cor* may be a favorable alternative over EFA in the following aspects:

1. *K-means sdm* shows more accurate results for the *item assignments* to factors, especially for small sample sizes
2. The two approaches do not involve model assumptions about the composition of the variance and covariance of the observed variables
3. Also with k-means clustering, it is possible to compute *cluster scores* when considering the mean of the cluster as the overlying construct and these *cluster scores* are determined

In fact, k-means clustering could overcome one of the problems of hierarchical CA while maintaining the advantage of CA over EFA that no crucial assumptions about an uncorrelated measurement error have to be made. This problem of hierarchical CA is that it is solely based on distances between items and therefore no statement is made about the construct to be measured. In k-means clustering the centroid of a cluster can be regarded as the overlying construct. Not only its coordinates are known but also distances between the construct and the items can be computed. Given that these distances are obtained directly from correlations, they can also directly be translated back into correlations. Accordingly, correlation matrices and even partial correlation matrices of items and constructs, equivalent to the loading matrix in PCA, can be obtained. Based on these distance measures or loadings again, the person scores on the constructs can be computed, the *cluster scores*. Just like in PCA and in contrary to PAF these *cluster scores* are determined since there is no residual variance involved in the equation. The equation of the linear combinations of construct scores is just the same as the one for component scores in PCA:

$$\mathbf{S} = \mathbf{XWX}'$$

where **S** is the n x m matrix of the construct scores, **X** is the n x p data matrix and **W** is the p x m weight matrix with the component loadings in case of PCA and the distances between items and the respective cluster means in the case of CA. This information might be of use for practitioners in psychological assessment. However, in this study we did not go deeper into the precise mathematics behind these calculations.

It shall be noted that the clustering method *k-means sdm* provides an additional advantage over other structure detection methods. The distances between the items and thus also the distances between items and the cluster mean doe not change when more items are added to the survey. This is not the case either for EFA or for other cluster techniques.
As for *dimensionality assessment*, EFA generally still performed as well as the combination *k-means cor* with Silhouette. This effect though was only shown in the Traditional simulation.

PA-PAF was sensitive to sample size and did therefore not perform so good in the Real World simulation where it indicated systematically more factors in the population data set than in the sub samples. However, when using PA-PAF for *dimensionality assessment* and preliminary assigning the items with the new k-means approach *k-means sdm*, the best fitting CFA model was found in the cross-validation with the NEO-PI-R data. For the IST-2000-R data PA-PCA outperformed PA-PAF. This result suggests that a combination of PA for *dimensionality assessment* and *k-means sdm* for the *item assignment* is most recommendable as exploratory structure detection method for practitioners especially for small sample sizes. To be more precise, PA-PCA shall be used when the aim is to reduce the dimensionality of the data to fewer underlying components that explain as much variance as possible. This combination moreover is useful when aiming to benefit from the absence of a measurement model including uncorrelated residual variances. Another option that still has to be investigated is combining *k-means cor* with silhouette for *dimensionality assessment* and *k-means sdm* for *item assignment*. In this study, there was no need to examine whether the new k-means approach can cope with the practical problems of EFA mentioned in the introduction, e.g. its difficulty with high cross loadings. By using parameters from real data we ensured to test practical meaningful conditions.

It should be stressed though, that CFA fit is only one criterion and is again based on the factor model. As mentioned above, a factor model is not necessary for CA, and therefore, in the future the evaluation of k-means approaches for clustering of items should focus more on the predictive quality of *cluster scores* obtained from its analysis.

Amthauer, R., Brocke, R., Liepmann, D., & Beauducel, A. (2007). *Intelligenz-Struktur-Test 2000 R - IST 2000-R* (2. erw.). Göttingen: Hogrefe.
Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.
Arrindell, W. A., & Van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, *9*(2), 165–178. Retrieved from http://apm.sagepub.com/content/9/2/165.short
Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438.
Bacon, D. R. (2001). An evaluation of cluster analytic approaches to initial model specification. *Structural Equation Modeling*, *8*(3), 397–429.
Borg, I., & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media.
Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). *clValid: An R Package for Cluster Validation* (Vol. 25).
Cattell, R. B. (1987). *IntelligenceIts Structure, Growth and Action: Its Structure, Growth and Action*. Elsevier.
Chavent, M., Genuer, R., Kuentz-Simonet, V., Liquet, B., & Saracco, J. (2013). ClustOfVar : an R package for dimension reduction via clustering of variables. Application in supervised classification and variable selection in gene expressions data. Presented at the Statistical Methods for (post)-Genomics Data, Amsterdam.
Chavent, M., Kuentz, V., Liquet, B., & Saracco, L. (2011). *ClustOfVar: An R Package for the Clustering of Variables*.
Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the problem of sample size: A clarification. *Psychological Bulletin*, *109*(3), 512–519.
Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272. Retrieved from http://psycnet.apa.org/journals/met/4/3/272/
Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265.
Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic dat analysis. *Bioinformatics*, *21*(15).
Horn, J. L. (1965a). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.
Horn, J. L. (1965b). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.
Hunter, J. E. (1973). Methods of reordering the correlation matrix to facilitate visual inspection and preliminary cluster analysis. *Journal of Educational Measurement*, *10*(1), 51–61.
Loevinger, J., Gleser, G. C., & DuBois, P. H. (1953). Maximizing the discriminating power of a multiple-score test. *Psychometrika*, *18*(4), 309–317.
MacCallum, R., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*(3), 502–511.
Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179.
Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15327574ijt0502_4
Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und*

McCrae, Revidierte Fassung. Göttingen: Hogrefe.
Patil, V. H., Singh, S. N., Mishra, S., & Todd Donavan, D. (2008). Efficient theory development and factor retention criteria: Abandon the "eigenvalue greater than one"criterion. *Journal of Business Research*, *61*(2), 162–170. Retrieved from http://www.sciencedirect.com/science/article/pii/S014829630700152X
Peterson, R. A. (2000). A Meta-Analysis of Variance Accounted for and Factor Loadings in Exploratory Factor Analysis. *Marketing Letters*, *11*(3), 261–275. doi:10.1023/A:1008191211004
Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846–850.
Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57–74. Retrieved from ´
Revelle, William. (2012). psych: Procedures for Personality and Psychological Research. (Version 1.0-91).
Sass, D. A. (2010). Factor loading estimation error and stability using exploratory factor analysis. *Educational and Psychological Measurement*, *70*(4), 557–577.
Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*(1), 73–103.
Schmitt, T. A. (2011). Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304–321. Retrieved from http://jpa.sagepub.com/content/29/4/304.short
Schonemann, P. H., & Steiger, J. H. (1978). On the validity of indeterminate factor scores. *Bulletin of the Psychonomic Society*, *12*(4), 287–290.
Schweizer, K. (1991). Classifying variables on the basis of disaggregate correlations. *Multivariate Behavioral Research*, *26*(3), 435–455.
Tryon, R. C. (1935). A theory of psychological components–an alternative to" mathematical factors.". *Psychological Review*, *42*(5), 425–454.
Tryon, R. C. (1939). *Cluster Analysis: Correlation Profile and Orthometric Factor Analysis for the Isolation of Unities in Mind and Personality*. Ann Arbor: Edward Brothers.
Tryon, R. C. (1959). Domain sampling formulation of cluster and factor analysis. *Psychometrika*, *24*(2), 113–135.
Tryon, Robert Choate, & Bailey, D. E. (1970). *Cluster analysis*. McGraw-Hill.
Van der Linden, D., Tsaousis, I., & Petrides, K. V. (2012). Overlap between General Factors of Personality in the Big Five, Giant Three, and trait emotional intelligence. *Personality and Individual Differences*, *53*(3), 175–179.
Van Dongen, S., & Enright, A. J. (2012). Metric distances derived from cosine similiarity and pearson and spearman correlations.
Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Coustruct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors of components. In *Problems and solutions in human assessment* (pp. 41–71). Boston: Kluwer.
Velicer, Wayne F. (1976). The relation between factor score estimates, image scores, and principal component scores. *Educational and Psychological Measurement*, *36*(1), 149–159.
Velicer, Wayne F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*(2), 231.
Vigneau, E., & Qannari, E. M. (2003). Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation*, *32*(4), 1131–1150.
Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432.