Computational Statistics and Data Analysis 54 (2010) 1858–1871

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Factorial and reduced K-means reconsidered

Marieke E. Timmerman [a,*], Eva Ceulemans [b], Henk A.L. Kiers [a], Maurizio Vichi [c]

[a] Heymans Institute of Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, The Netherlands

[b] Department of Educational Sciences, Katholieke Universiteit Leuven, Andreas Vesaliusstraat 2, B-3000 Leuven, Belgium

[c] Dipartimento di Statistica, Probabilità e Statistiche Applicate, Sapienza Università di Roma, P. le A. Moro 5, I-00185 Rome, Italy

ARTICLE INFO

Article history:
Received 8 September 2008
Received in revised form 12 February 2010
Accepted 13 February 2010
Available online 19 February 2010

Keywords:
Cluster analysis
Factorial model
K-means
Dimensional reduction

ABSTRACT

Factorial K-means analysis (FKM) and Reduced K-means analysis (RKM) are clustering methods that aim at simultaneously achieving a clustering of the objects and a dimension reduction of the variables. Because a comprehensive comparison between FKM and RKM is lacking in the literature so far, a theoretical and simulation-based comparison between FKM and RKM is provided. It is shown theoretically how FKM's versus RKM's performances are affected by the presence of residuals within the clustering subspace and/or within its orthocomplement in the observed data. The simulation study confirmed that for both FKM and RKM, the cluster membership recovery generally deteriorates with increasing amount of overlap between clusters. Furthermore, the conjectures were confirmed that for FKM the subspace recovery deteriorates with increasing relative sizes of subspace residuals compared to the complement residuals, and that the reverse holds for RKM. As such, FKM and RKM complement each other. When the majority of the variables reflect the clustering structure, and/or standardized variables are being analyzed, RKM can be expected to perform reasonably well. However, because both RKM and FKM may suffer from subspace and membership recovery problems, it is essential to critically evaluate their solutions on the basis of the content of the clustering problem at hand.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Cluster analysis aims at assigning a number of objects to a limited number of homogeneous classes. This is often done on the basis of the objects' scores on multiple variables. The inclusion of variables in a cluster analysis that hardly reflect, or even do not reflect the clustering structure may hinder or even completely obscure the recovery of the underlying cluster structure (e.g. Milligan, 1996). To deal with those problems, various approaches can be taken. One may use variable selection (e.g. Steinley and Brusco, 2008) or variable weighting (e.g. Milligan and Cooper, 1988), implying that some variables are discarded from the cluster analysis or are given less weight. An alternative is the subspace clustering approach, which rests on the assumption that the cluster centroids are located in a subspace of the variables.

A relatively easy subspace clustering approach, which does not require distributional assumptions, uses component analysis. The first attempt in this direction was a two-step procedure, where a principal component analysis was followed by a cluster analysis. After various warnings against this so-called tandem clustering (e.g. Arabie and Hubert, 1994), De Soete and Carroll (1994) proposed Reduced K-means analysis (RKM), which appeared to equal the earlier proposed Projection Pursuit Clustering (Bock, 1987). RKM simultaneously searches for a clustering of the objects, based on the K-means criterion (MacQueen, 1967), and a dimension reduction of the variables, based on component analysis. The notion that RKM may fail

* Corresponding author. Tel.: +31 50 3636255; fax: +31 50 3636304.
E-mail address: m.e.timmerman@rug.nl (M.E. Timmerman).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.02.009

to recover the clustering of the objects when the data contain much variance in directions orthogonal to the subspace of the data in which the clusters reside, led to the proposal of Factorial K-means analysis (FKM Vichi and Kiers, 2001).

A comprehensive comparison between FKM and RKM is lacking, both theoretically and empirically. After all, the simulation study that Vichi and Kiers (2001) conducted to support their claims about the superior performance of FKM over RKM consisted of the analysis of a single simulated data set only. This means that it is unknown when RKM and/or FKM would yield good insight into the cluster structure in empirical data, and hence when to use RKM or FKM. The purpose of the present article is to clarify this issue, so that FKM and RKM can be sensibly used in practice. It will be shown that FKM and RKM serve different goals. Therefore FKM and RKM complement each other: FKM is of use when RKM fails, and vice versa.

The remainder of the paper is organized as follows. Section 2 recapitulates the RKM and FKM models. Section 3 provides a theoretical comparison of the performance of RKM and FKM. Section 4 presents a simulation study to examine the conjectures from Section 3, and to evaluate the performance of RKM and FKM in various conditions. The use of FKM and RKM is illustrated with an empirical example in Section 5. The paper closes with a discussion of the reported findings (Section 6).

2. Factorial and reduced K-means

2.1. Notation

The following notation is adopted in this paper:

I	Number of objects, indexed $i = 1, \ldots, I$.
J	Number of variables, indexed $j = 1, \ldots, J$.
C	Number of clusters, indexed $c = 1, \ldots, C$.
Q	Number of components to which the variables are reduced, indexed $q = 1, \ldots, Q$.
$\mathbf{X}$	An $I \times J$ matrix containing the observed scores of I objects on J variables. Variables are supposed to be centered. In the case they have different units of measurements they are commonly standardized, i.e., so that they have a mean of zero and unit variances.
$\mathbf{U}$	A binary $I \times C$ membership matrix, which specifies to which cluster each object belongs, i.e., $u_{ic} = 1$ if object i belongs to cluster c, and $u_{ic} = 0$ otherwise; $\sum_{c=1}^{C} u_{ic} = 1$.
$\mathbf{F}$	$C \times Q$ centroid matrix, where f_{cq} is the centroid score of cluster c on component q.
$\mathbf{A}$	$J \times Q$ columnwise orthonormal loading matrix; i.e., $\mathbf{A}'\mathbf{A} = \mathbf{I}_Q$.
$\mathbf{A}^{\perp}$	$J \times (J - Q)$ columnwise orthonormal matrix for which it holds that $\mathbf{A}'\mathbf{A}^{\perp} = \mathbf{0}$.
$\mathbf{E}\mathbf{A}'$	$I \times Q$ subspace residual matrix.
$\mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}$	$I \times (J - Q)$ complement residual matrix.
$\mathbf{T}$	$Q \times Q$ orthonormal rotation matrix.
$\mathbf{0}_{I\times J}$	$I \times J$ matrix consisting of zeros.
$\mathbf{I}_J$	$J \times J$ identity matrix.
diag($\mathbf{c}$)	Diagonal matrix with diagonal elements equal to the elements of vector $\mathbf{c}$.

2.2. Loss functions and decomposition rules

Both FKM and RKM aim at simultaneously finding an optimal partitioning of the objects and an optimal reduction of the variables. As such, FKM as well as RKM decompose the data matrix $\mathbf{X}$ into a membership matrix $\mathbf{U}$, a columnwise orthonormal loading matrix $\mathbf{A}$ that reveals the extent to which the variables express the clustering structure, and a centroid matrix $\mathbf{F}$ that contains the scores of the cluster centroids on the Q components. To illustrate the interpretation of these matrices, we make use of the following artificial example, with $I = 4, J = 5, C = 2$, and $Q = 2$:

$$\mathbf{U} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{F} = \begin{bmatrix} 1.2 & 1.3 \\ -0.1 & 0.2 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 0.7 & 0.0 \\ 0.7 & 0.0 \\ 0.0 & 0.7 \\ 0.0 & 0.7 \\ 0.0 & 0.0 \end{bmatrix}.$$

The membership matrix $\mathbf{U}$ shows that Objects 1 and 2 belong to one cluster, and Objects 3 and 4 to another. The centroid matrix contains the centroids of the two clusters, in a two-dimensional space. The loading matrix $\mathbf{A}$ shows that Variables 1 and 2 are associated to the first dimension, Variables 3 and 4 to the second, and Variable 5 does not reflect the clustering structure at all. Such a variable is commonly denoted as a noise or masking variable. Note that, because the size of loadings express the relative importance of the variables concerned in the clustering, FKM and RKM are related to variable selection and weighting approaches. The essential difference is that in the latter two approaches the variables are selected and weighted before cluster analysis, whereas FKM and RKM weight, select and cluster simultaneously.

The key difference between FKM and RKM can be easily seen in the objective functions that are associated with these models. The FKM loss function to minimize is

$$F_{\mathrm{FKM}}(\mathbf{U},\mathbf{F},\mathbf{A}) = \|\mathbf{XAA}' - \mathbf{UFA}'\|^2 = \|\mathbf{XA} - \mathbf{UF}\|^2, \tag{1}$$

whereas the RKM loss function to minimize can be written as

$$F_{\mathrm{RKM}}(\mathbf{U},\mathbf{F},\mathbf{A}) = \|\mathbf{X} - \mathbf{UFA}'\|^2. \tag{2}$$

Note that when the centroids are located in the full space (i.e., $Q = J$), both FKM and RKM equal the well-known K-means clustering. Note further that FKM as well as RKM have rotational indeterminacy for the loadings and the centroid scores, and permutational indeterminacy for the clusters. This implies that without altering the value of the loss function the loading matrix $\mathbf{A}$ can be replaced by $\mathbf{AT}$, where $\mathbf{T}$ is a $(Q \times Q)$ orthonormal rotation matrix, provided that this rotation is compensated for in the centroid matrix as $\mathbf{FT}$, and matrix $\mathbf{U}$ may be replaced by $\mathbf{U\Pi}$, where $\mathbf{\Pi}$ is a $(C \times C)$ permutation matrix, provided that the permutation is compensated for in the centroid matrix as $\mathbf{\Pi}'\mathbf{F}$.

As can be seen in the FKM loss function (1), FKM *minimizes* the sum of the squared distances between the centroids in the projected space and the projected data points (i.e., the observed data points that are projected onto the subspace in which the centroids reside). This is the within-clusters deviance in the reduced space. The RKM loss function (2) shows that RKM searches a partitioning of the objects that minimizes the sum of the squared distances between the observed data and the 'quasi' centroids, that is centroids that are located in a subspace of the data which is spanned by the columns of $\mathbf{A}$.

2.3. Subspace residuals and complement residuals

To further compare the RKM and FKM analyses, it is instructive to examine the models that are being fitted by (1) and (2). The RKM model that is being fitted by (2) is simply

$$\mathbf{X} = \mathbf{UFA}' + \mathbf{E}_R, \tag{3}$$

where $\mathbf{E}_R$ denotes an $(I \times J)$ residual matrix. The optimal $\mathbf{F}$ is given by $\mathbf{F} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}$, thus model (3) can be rewritten as

$$\mathbf{X} = \mathbf{H}_U\mathbf{XAA}' + \mathbf{E}_R, \tag{4}$$

where $\mathbf{H}_U = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$ is the projection matrix on the space spanned by the columns of $\mathbf{U}$. Note that $(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'(C \times I)$ indicates how the different objects are weighted when computing the C centroids.

The FKM model, as specified by Vichi and Kiers (2001, Equation 1), is

$$\mathbf{XAA}' = \mathbf{UFA}' + \mathbf{E}_{\mathrm{F}} \tag{5}$$

with residual matrix $\mathbf{E}_{\mathrm{F}}(I \times J)$, which can be rewritten by including the optimal $\mathbf{F}$ as

$$\mathbf{XAA}' = \mathbf{H}_U\mathbf{XAA}' + \mathbf{E}_{\mathrm{F}}. \tag{6}$$

The RKM model reconstructs all the data with only the centroids lying in the reduced space, while FKM assumes that both the centroids and objects lie in reduced space.

The difference between FKM and RKM can be clearly expressed in terms of their residuals. We start our discussion by considering the residuals in the FKM model. Because $\mathbf{E}_{\mathrm{F}} = \mathbf{XAA}' - \mathbf{UFA}'$, the matrix $\mathbf{E}_{\mathrm{F}}$ lies in the row space of $\mathbf{A}'$. This implies that the residual matrix $\mathbf{E}_{\mathrm{F}}$ can be written as $\mathbf{E}_{\mathrm{F}} = \mathbf{EA}'$, where the matrix $\mathbf{E}$ is in the same column space as $\mathbf{E}_{\mathrm{F}}$. Thus, $\mathbf{E}_{\mathrm{F}}$ contains the subspace residuals of the projected data when represented in the full space of $\mathbf{X}$. Using that $\mathbf{XAA}' = \mathbf{UFA}' + \mathbf{E}_{\mathrm{F}} = \mathbf{UFA}' + \mathbf{EA}' = (\mathbf{UF} + \mathbf{E})\mathbf{A}'$, and recalling that the rotated projected configuration of objects does not alter their squared Euclidean distances from the rotated centroids, the FKM model in (5) can be rewritten as $\mathbf{XA} = \mathbf{UF} + \mathbf{E}$. Thus, FKM offers a model for the orthogonal projection onto a subspace of the data, rather than for the data themselves. To model the observed data exhaustively (i.e., to model $\mathbf{X}$ rather than $\mathbf{XAA}'$), another residual term is needed, describing residuals in the orthocomplement of $\mathbf{A}$, i.e., in the orthogonal complement subspace of $\mathbf{A}$. Those "complement residuals", which equal $\mathbf{X} - \mathbf{XAA}'$, are denoted here by $\mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}$, where $\mathbf{A}^{\perp}$ is a $J \times (J - Q)$ columnwise orthonormal matrix for which it holds that $\mathbf{A}'\mathbf{A}^{\perp} = \mathbf{0}$. The full model for the observed data $\mathbf{X}$ can then be described as

$$\mathbf{X} = \mathbf{UFA}' + \mathbf{EA}' + \mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}, \tag{7}$$

which forms the basis for our discussion of the performance of FKM and RKM in the remainder of this paper.

Formula (7) implies that the scores of I objects on J variables consist of a structural part and two types of residuals, namely $\mathbf{EA}'$, which are residuals within the subspace where centroids or centroids and objects lie (for RKM and FKM, respectively), and $\mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}$, which are residuals within the complement of this subspace; from now on those residuals will be called subspace residuals and complement residuals, respectively.

From (7) it is clear that a distinction can be made between 4 types of data: data that contain neither subspace nor complement residuals (i.e., $\mathbf{E} = \mathbf{0}$ and $\mathbf{E}^{\perp} = \mathbf{0}$), data that contain subspace residuals only, data that contain complement residuals only, and, finally, data that contain both subspace and complement residuals. Simple examples of these four types of data are given in Fig. 1, in which variables $x1$ and $x2$ are observed variables (i.e., $J = 2$), the number of clusters C equals 2, and the number of components Q equals 1.

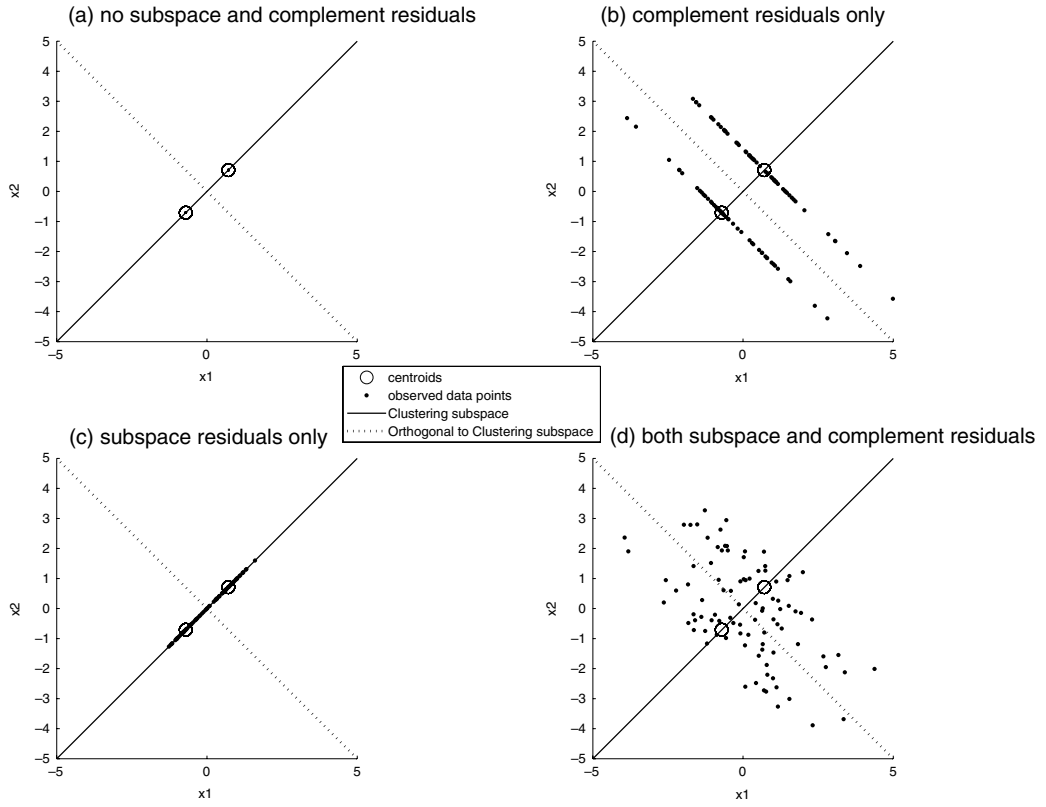

Fig. 1. Examples of data with (a) neither subspace nor complement residuals; (b) complement residuals only; (c) subspace residuals only; (d) both subspace and complement residuals; note that the subspace chosen here is the one dimensional subspace represented by the diagonal line from bottom left to top right.

The decomposition of the data into a structural part and two types of residual parts offers further insight into the nature of masking variables. Masking variables do not reflect the underlying clustering structure, and therefore are assumed to be represented fully in the residual part of the model. Eq. (7) together with this assumption implies that for masking variables the corresponding rows in $\mathbf{A}$ (referring to the weights involved in obtaining the variables as linear combinations of the underlying variables) must be zero; indeed, if they were not, such variables would contain, at least to some extent, the clustering structure in $\mathbf{U}$. As a consequence, masking variables are fully expressed by the third term, that is $\mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}$. In other words, the scores on the masking variable equal the complement residuals for that masking variable.

2.4. Perfect FKM data and perfect RKM data

Before we go on to a theoretical comparison of the RKM and FKM performances, it is instructive to consider the class of observed data that perfectly comply with FKM and RKM, that is, for which the FKM and RKM loss functions are 0. From (1) it follows that the FKM loss function is 0 if and only if $\mathbf{XA} = \mathbf{UF}$. To specify the class of perfect FKM data, we first note that every matrix $\mathbf{X}(I \times J)$ can be expressed in terms of $\mathbf{A}$ (a $(J \times Q)$ columnwise orthonormal loading matrix) and $\mathbf{A}^{\perp}$ as

$$\mathbf{X} = \mathbf{BA}' + \mathbf{CA}^{\perp\prime}, \tag{8}$$

for certain matrices $\mathbf{B}$ $(I \times Q)$ and $\mathbf{C}$ $(I \times (J - Q))$. Inserting (8) for $\mathbf{X}$ in $\mathbf{XA} = \mathbf{UF}$, we find $\mathbf{XA} = \mathbf{BA}'\mathbf{A} + \mathbf{CA}^{\perp\prime}\mathbf{A} = \mathbf{B} = \mathbf{UF}$. Thus, for all matrices $\mathbf{X}$ that contain perfect FKM data, we have $\mathbf{X} = \mathbf{UFA}' + \mathbf{CA}^{\perp\prime}$. It is also readily verified that such data satisfy the FKM model for any matrix $\mathbf{C}$. Denoting this arbitrary matrix $\mathbf{C}$ as $\mathbf{E}^{\perp}$, we can describe the full class of perfect FKM data as

$$\mathbf{X} = \mathbf{UFA}' + \mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}. \tag{9}$$

Comparing this expression to that in (3), we see that perfect FKM data are characterized by data in which the subspace residuals are 0, whereas the complement residuals need not be 0.

From (2) it follows that the RKM loss function is 0 if and only if $\mathbf{X} = \mathbf{UFA}'$. Thus, perfect RKM data are those data for which both the subspace residuals and the complement residuals are 0. Therefore, perfect RKM data are perfect FKM data, but not vice versa. This is illustrated in Fig. 1, where perfect FKM data are shown in Fig. 1(a) and (b), and perfect RKM data in Fig. 1(a) only.

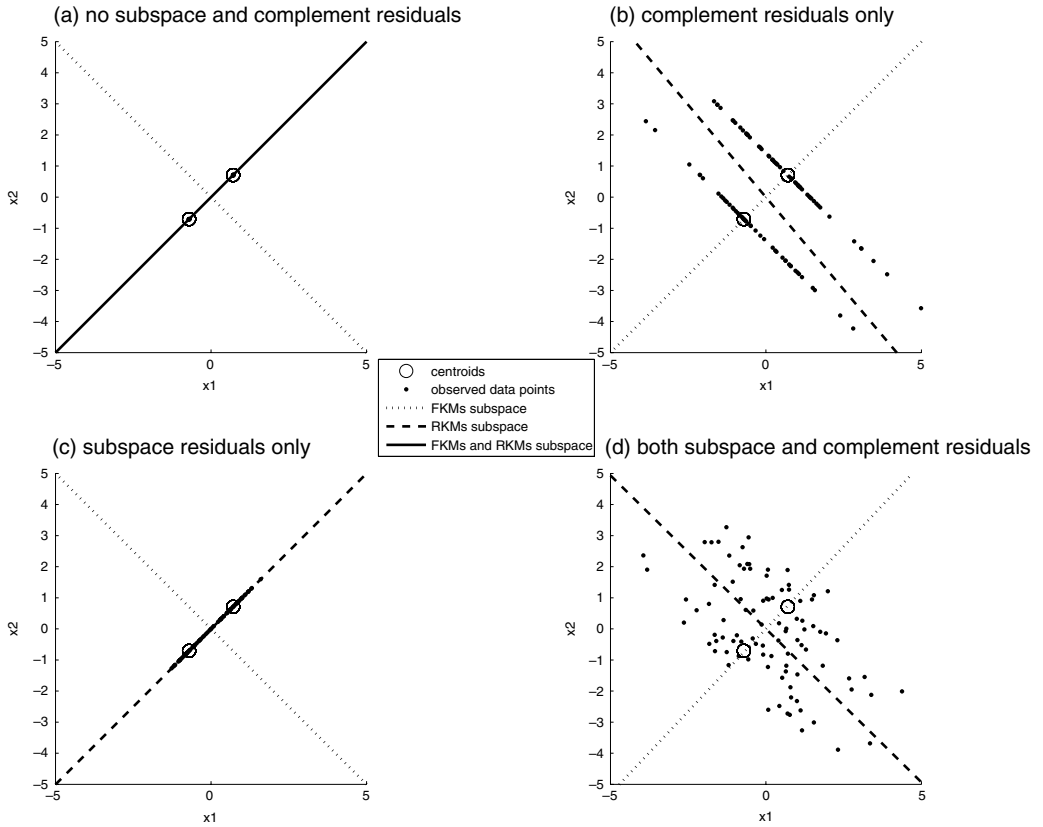

Fig. 2. Examples of possible FKM and RKM solutions for data with (a) neither subspace nor complement residuals; (b) complement residuals only; (c) subspace residuals only; (d) both subspace and complement residuals. The lines in the figures indicate the subspace(s) where RKM and/or FKM may end up.

3. Theoretical comparison: FKM's versus RKM's performances

In view of the equal objectives and rather similar model specifications of RKM and FKM, it is important to understand when either or both of the methods would perform well in recovering a cluster structure. We successively examine FKM's and RKM's performances in the cases of the four different types of data that were distinguished in Section 2 and illustrated in Fig. 1. The possible FKM and RKM solutions for the data in Fig. 1 are presented in Fig. 2.

3.1. Data with zero subspace residuals and zero complement residuals

First, we consider data with zero subspace residuals and zero complement residuals, that is, data that comply perfectly with the FKM as well as the RKM model (Fig. 1(a)). It will turn out that in this case, RKM yields a perfect solution, whereas FKM may easily yield arbitrary solutions that, however, are easily detectable.

The key problem for FKM in the case of data without any residuals is that its loss function is 0 for all possible projections of the observed data, rather than only the projection on the subspace with the clustering. This can be seen as follows: Suppose for all practical purposes that the complement space has more dimensions than the subspace of interest. A projection on a subspace orthogonal to the one with the clustering structure implies that the estimated loading matrix $\hat{\mathbf{A}}$ is $\mathbf{A}^{\perp}\mathbf{T}$, with $\mathbf{T}$ an arbitrary orthonormal matrix. As a result, the projected data coordinates equal $\mathbf{X}\hat{\mathbf{A}} = \mathbf{U}\mathbf{F}\mathbf{A}'\mathbf{A}^{\perp}\mathbf{T} = \mathbf{U}\mathbf{F}\mathbf{0} = \mathbf{0}$. Thus, the loss function value will be zero by taking an arbitrary $\hat{\mathbf{U}}$ and $\hat{\mathbf{F}} = \mathbf{0}$. Of course, the FKM loss function is also zero by taking $\hat{\mathbf{A}}$ within the subspace itself, hence as $\mathbf{AT}$, for any orthonormal matrix $\mathbf{T}$. Hence, in the absence of residuals, the minimum value of the FKM loss function is zero, irrespective whether the estimated loading matrix is in the column space of $\mathbf{A}$ and/or in its orthocomplement $\mathbf{A}^{\perp}$. Note that estimates residing purely in the orthocomplement subspace can be easily diagnosed, because the total deviance of $\mathbf{X}\hat{\mathbf{A}}$ is zero. Note further that $\hat{\mathbf{A}}$ can only equal $\mathbf{A}^{\perp}\mathbf{T}$ when $Q \leq (J - Q)$.

In contrast to FKM, the RKM loss function is zero if and only if $\hat{\mathbf{A}}$ is $\mathbf{AT}$ for some orthonormal matrix $\mathbf{T}$, $\hat{\mathbf{U}}$ and $\mathbf{U}$ differ at most by a permutation. That is, $\hat{\mathbf{U}} = \mathbf{U}\mathbf{\Pi}$, for some permutation matrix $\mathbf{\Pi}$, and $\hat{\mathbf{F}} = \mathbf{\Pi}'\mathbf{F}\mathbf{T}$, provided that $\text{rank}(\mathbf{U}) = C$ and $\mathbf{F}$ has rank $Q (Q \leq C)$ and does not have any equal rows, and $\text{rank}(\mathbf{A}) = Q$, $Q \leq \min(C, J)$, as has been proven in Appendix A. This implies that RKM yields a perfect solution in case of zero residuals. (Note that because $\mathbf{F}\mathbf{A}'$ cannot be fit better by using more than C components, we can assume $Q \leq C$ without loss of generality.)

In Fig. 2(a), it is illustrated that in the absence of residuals, RKM yields a correct estimate of the subspace, whereas FKM may end up in two solutions of which only one is correct.

3.2. Data with complement residuals only

Secondly, we consider perfect FKM data (see (9)), with complement residuals of full rank $J-Q$. We only consider the case where the estimated matrices have the same orders as the underlying matrices, i.e., the correct numbers of components and clusters are being estimated.

Now suppose that $\text{rank}(\mathbf{U}) = C$, $\text{rank}([\mathbf{UF}|\mathbf{E}^{\perp}]) = (Q+J)$, where $[\mathbf{UF}|\mathbf{E}^{\perp}]$ is a block matrix consisting of matrices $\mathbf{UF}$ and $\mathbf{E}^{\perp}$ positioned next to each other, $\text{rank}(\mathbf{A}) = Q$, $Q \leq J$, $\mathbf{F}$ does not have any equal rows, and $\mathbf{E}^{\perp}$ has no clustering structure whatsoever (which are, all in all, realistic conditions). If those conditions hold, it can be proven that, as in the above case, the FKM estimates are as follows: $\hat{\mathbf{A}}$ is $\mathbf{AT}$, $\hat{\mathbf{U}}$ and $\mathbf{U}$ differ at most by a permutation, i.e., $\hat{\mathbf{U}} = \mathbf{U\Pi}$, for some permutation matrix $\mathbf{\Pi}$, and $\hat{\mathbf{F}} = \mathbf{\Pi}'\mathbf{FT}$. This proof is provided in Appendix B.

For data with complement residuals only, the RKM loss function boils down to

$$\begin{aligned} g_1(\mathbf{U},\mathbf{F},\mathbf{A}) &= \left\| \left(\mathbf{UFA}' + \mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}\right) - \hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}' \right\|^2 \\ &= \left\| \mathbf{UFA}' - \hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}' \right\|^2 + \left\| \mathbf{E}^{\perp}\mathbf{A}^{\perp\prime} \right\|^2 - 2\text{tr}\left(\hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}'\mathbf{A}^{\perp}\mathbf{E}^{\perp\prime}\right) + 2\text{tr}\left(\mathbf{UFA}'\mathbf{A}^{\perp}\mathbf{E}^{\perp\prime}\right) \\ &= \left\| \mathbf{UFA}' - \hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}' \right\|^2 + \left\| \mathbf{E}^{\perp}\mathbf{A}^{\perp\prime} \right\|^2 - 2\text{tr}\left(\hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}'\mathbf{A}^{\perp}\mathbf{E}^{\perp\prime}\right). \end{aligned} \tag{10}$$

Clearly, even when RKM's estimates of $\mathbf{U}$, $\mathbf{F}$, and $\mathbf{A}$ are perfect, the loss function will not be zero but $\|\mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}\|^2$. Hence, RKM may end up in a solution which resides partly in the subspace defined by $\mathbf{A}^{\perp}$ if this solution has a lower loss function value. An example of the latter situation is presented in Fig. 2(b).

3.3. Data with subpace residuals only

Thirdly, we consider observed data with subspace residuals present but no complement residuals (i.e., $\mathbf{X} = \mathbf{UFA}' + \mathbf{EA}'$). In this case, FKM will always provide an arbitrary estimate of the membership matrix, as can be seen as follows. When the estimated loading matrix $\hat{\mathbf{A}}$ would equal $\mathbf{A}^{\perp}\mathbf{T}$, the projected data are zero (because $\mathbf{XA}^{\perp}\mathbf{T} = (\mathbf{UFA}' + \mathbf{EA}')\mathbf{A}^{\perp}\mathbf{T} = \mathbf{0}$). This implies that a zero loss value is obtained with zero estimates of $\mathbf{F}$, and arbitrary estimates of $\mathbf{U}$. The FKM loss value with an estimated loading matrix $\hat{\mathbf{A}}$ that fully or partly resides in the subspace spanned by $\mathbf{A}$ will result in a loss value larger than zero, and hence solutions with $\hat{\mathbf{A}} = \mathbf{A}^{\perp}\mathbf{T}$ are favoured. Note that this solution is clearly trivial, because the total deviance of $\mathbf{XA}^{\perp}$ equals zero.

For observed data containing subspace residuals only, the RKM loss value will always be larger than zero. The subspace solution will be the correct one (because any part of the solution in $\mathbf{A}^{\perp}$ would add to the loss function), but its performance in recovering the clustering structure is for obvious reasons expected to decrease with increasing subspace residual variances.

An illustration of the performance of RKM and FKM for data with subspace residuals only can be found in Fig. 2(c).

3.4. Data with subspace as well as complement residuals

Finally, we consider observed data with residuals both in the clustering subspace and in the subspace orthogonal to the clustering subspace. This implies that the observed data can be written as

$$\mathbf{X} = \mathbf{UFA}' + \mathbf{EA}' + \mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}. \tag{11}$$

Note that $\mathbf{EA}'$ denotes the subspace residuals in the full space and $\mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}$ the complement residuals. In this case, we cannot make any hard statements on the possible solutions resulting from FKM or RKM. Therefore, we will study what happens in such cases by means of a simulation study. We do have some conjectures, however, which are based on the results with only one type of residuals present in the data. That is, we have seen that in the case with only subspace residuals, RKM ends up in the correct subspace, while FKM does not. The reverse holds in the case of only complement residuals. We conjecture that similar results hold when one of the two types of residuals is generally much smaller than the other type of residuals. Specifically, we conjecture that, if subspace residuals are much larger than complement residuals, then RKM can be expected to outperform FKM, and the reverse can be expected if subspace residuals are much smaller than complement residuals.

4. Simulation study

To gain insight into the correctness of our conjectures and into the comparative performances of RKM and FKM in the presence of different levels of complement residuals and subspace residuals, we conduct a simulation study. It is investigated to what extent RKM and FKM are capable of recovering the correct subspace and the correct cluster memberships in various conditions.

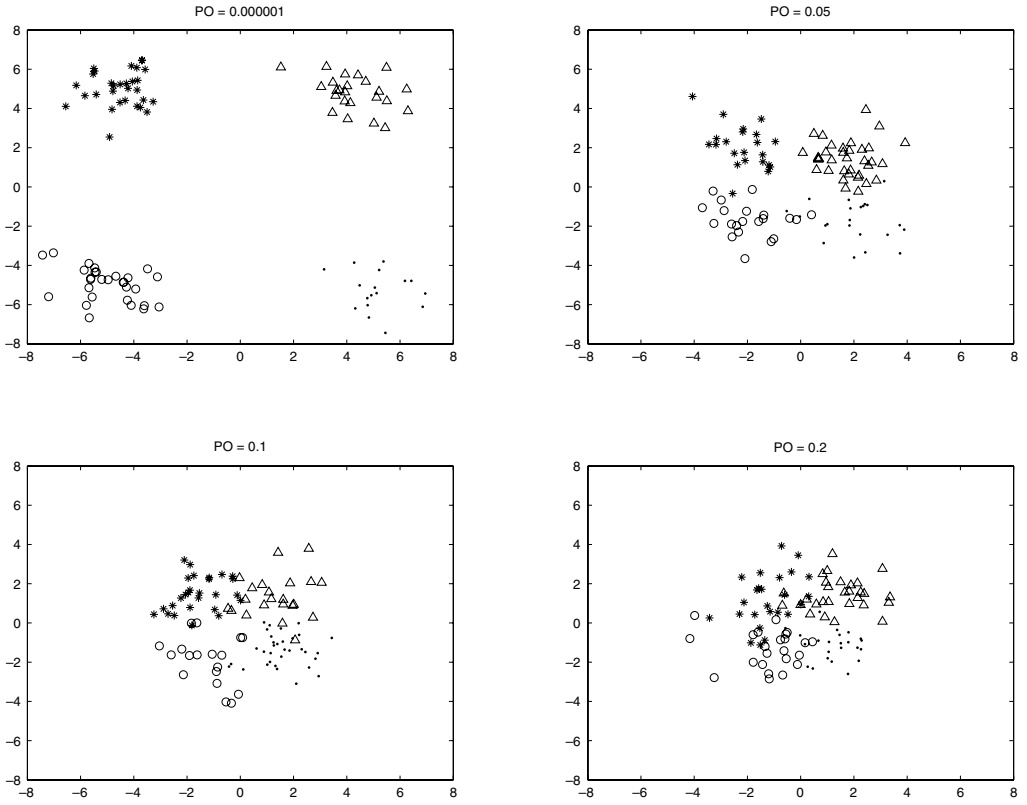

Fig. 3. Example of simulated object scores for 100 objects in 4 clusters in the correct 2-dimensional subspace at 4 levels of Proportion of Overlap (PO).

With respect to the recovery of the subspace it was conjectured that with increasing relative sizes of subspace residuals compared to the complement residuals the recovery for FKM would deteriorate, whereas the recovery for RKM would ameliorate. For the recovery of the cluster memberships, it was expected for both FKM and RKM that for those cases for which the correct subspace would be estimated, the recovery of the memberships would deteriorate when the clusters overlap more.

4.1. Generation of the simulated data and experimental design

Each simulated sample data matrix $\mathbf{X}_{\text{sim}}$ was generated as

$$\mathbf{X}_{\text{sim}} = \mathbf{U}_{\text{sim}}\mathbf{F}_{\text{sim}}\mathbf{A}'_{\text{sim}} + \mathbf{E}_{\text{sim}}\mathbf{A}'_{\text{sim}} + \mathbf{E}^{\perp}_{\text{sim}}\mathbf{A}^{\perp\prime}_{\text{sim}}. \tag{12}$$

In the simulation experiment, we fixed the expected proportion of objects in $\mathbf{U}_{\text{sim}}$ to be equal across clusters; the relative distances between the centroids were fixed by taking $\mathbf{F}_{\text{sim}} = [-c\ c]'$ and $\mathbf{F}_{\text{sim}} = \begin{bmatrix} c & -c & c & -c \\ c & -c & -c & c \end{bmatrix}'$, for the conditions with $Q = 1$ and $Q = 2$, respectively, where c denotes a scalar to manipulate the expected Proportion of Overlap between clusters; the distributions of the complement and subspace residuals were taken as $\mathbf{E}_{\text{sim}} \sim N(0, \sigma^2_{\mathbf{E}_{\text{sim}}} = 1)$ and $\mathbf{E}^{\perp}_{\text{sim}} \sim N(0, \sigma^2_{\mathbf{E}_{\text{sim}}{}^{\perp}})$. The orthonormal loading matrix was taken fixed as $\mathbf{A}_{\text{sim}} = [\,0.45 \quad 0.45 \quad 0.45 \quad 0.45 \quad 0.22 \quad 0.22 \quad 0.22 \quad 0.22\,]'$ for $Q = 1$ and $\mathbf{A}_{\text{sim}} = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 & 0.25 & 0.25 & 0.56 & 0.56 \\ 0.41 & 0.41 & 0.41 & -0.41 & -0.41 & -0.41 & 0.00 & 0.00 \end{bmatrix}'$ for $Q = 2$, implying that the number of observed variables equals 8; those matrices were chosen such that the loading matrix had a certain, but not fully simple structure, and that loadings varied at two levels across the variables. The matrix $\mathbf{A}^{\perp}_{\text{sim}}$ was taken as an orthonormal basis for the null space of $\mathbf{A}'_{\text{sim}}$, so that $\mathbf{A}'_{\text{sim}}\mathbf{A}^{\perp}_{\text{sim}} = \mathbf{0}$.

The following four factors were manipulated in the experiment:

1. The expected Proportion of Overlap between clusters in the correct subspace (PO) was varied by manipulating the distances between the centroids in $\mathbf{F}_{\text{sim}}$. PO was defined as the proportion of shared density between clusters, as proposed by Steinley and Henson (2005). PO was varied at 5 levels: 0.000001, 0.05, 0.10, 0.15 and 0.20.

To offer an impression of the effect of the manipulation of the Proportion of Overlap, an example of simulated object scores for 100 objects in 4 clusters in the correct 2-dimensional subspace (i.e., $\mathbf{U}_{\text{sim}}\mathbf{F}_{\text{sim}} + \mathbf{E}_{\text{sim}}$) is depicted in Fig. 3, for different levels of PO.

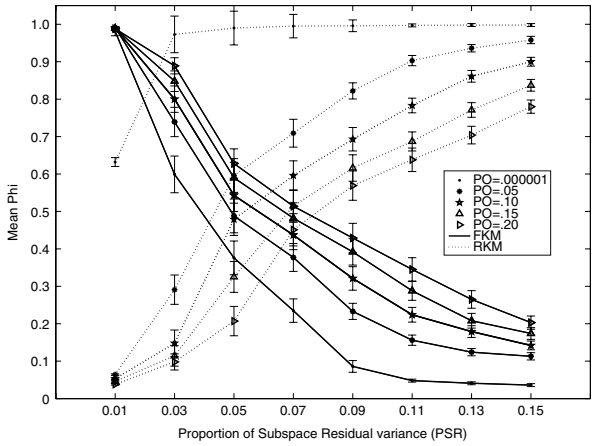

Fig. 4. Means and associated 95% Confidence Intervals of Phi coefficients as a function of Proportion of Subspace Residual variance (PSR) and Proportion of Overlap (PO).

2. The relative size of the variance of the subspace residuals and complement residuals was varied by manipulating the expected Proportion of Subspace Residual variance relative to the total residual variance ($\text{PSR} = \frac{\sigma^2_{\mathbf{E}_{\text{sim}}}}{\sigma^2_{\mathbf{E}_{\text{sim}}} + \sigma^2_{\mathbf{E}^{\perp}_{\text{sim}}}} = \frac{1}{1+\sigma^2_{\mathbf{E}^{\perp}_{\text{sim}}}}$, because $\sigma^2_{\mathbf{E}_{\text{sim}}}$ was taken fixed at 1). PSR was varied at 8 levels: 0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.13 and 0.15.

3. The Model Complexity (MC) was varied at 2 levels: $(Q = 1, C = 2)$ and $(Q = 2, C = 4)$.

4. The Number of Objects (NO) was varied at $NO = 50$, 100 and 200.

The experimental design was fully crossed, with 50 replicates per cell, yielding $5 \times 8 \times 3 \times 2 \times 50 = 12{,}000$ simulated data sets.

4.2. Analyses of the simulated data

All 12,000 simulated data sets were analyzed with both RKM and FKM in the correct complexity, that is, given the true values of C and Q. The alternating least squares algorithms used were taken from De Soete and Carroll (1994) for RKM, and from Vichi and Kiers (2001) for FKM. To decrease the chance of missing the global optimum, we used 1000 randomly started runs for each analysis.

4.3. Quality criteria

In this simulation study, it was studied how well RKM and FKM succeeded in recovering the correct subspace and the correct cluster memberships. The subspace recovery was assessed by considering the mean of the Phi coefficients (Tucker, 1951) between the columns of the estimated and simulated loading matrices ($\hat{\mathbf{A}}$ and $\mathbf{A}_{\text{sim}}$, respectively), where $\hat{\mathbf{A}}$ had been orthogonally Procrustes rotated (Cliff, 1966) towards $\mathbf{A}_{\text{sim}}$ to deal with the rotational indeterminacy. The Phi coefficient indicates to what extent two columns are proportional, and ranges from -1 to 1, where a value of $(-)1$ indicates perfect proportionality and a value of 0 no proportional relationship.

The cluster membership recovery was assessed by the Adjusted Rand Index (Hubert and Arabie, 1985). The Adjusted Rand Index has the maximal value of 1 in the case of a perfect recovery of the underlying clustering structure, and a value of 0 in the case where the membership matrices $\mathbf{U}_{\text{sim}}$ and $\hat{\mathbf{U}}$ do not correspond more than expected by chance.

4.4. Results

The quality of the Recoveries of the Subspace and the Cluster memberships will be discussed consecutively. Firstly, the simulation results are presented in view of the stated conjectures. Subsequently, factors that appeared to strongly influence the quality of recovery will be briefly discussed; those factors are identified by inspecting partial η^2-values resulting from a full-factorial Repeated Measures ANalysis Of VAriance (RMANOVA).

Recovery of the subspace

To obtain an impression of the quality of subspace recovery with RKM and FKM, Fig. 4 presents the means (and associated 95% CIs) of the Phi coefficients as a function of Proportion of Subspace Residual variance (PSR) and Proportion of Overlap (PO).

As can be seen in Fig. 4, the quality of subspace recovery decreases with increasing PSR for FKM, whereas for RKM the reverse holds. Those findings corroborate the hypothesis that, with increasing relative sizes of subspace residuals compared to the complement residuals, the subspace recovery would deteriorate for FKM, and ameliorate for RKM.

Table 1
Values of partial η^2 from full-factorial RMANOVAs on Phi coefficients and Adjusted Rand Indices, respectively.

Source	Phi coefficient	Adjusted Rand Index
Method	0.563	0.679
Method × PSR	0.917	0.873
Method × MC	0.788	0.705
Method × PO	0.726	0.701
Method × MC × PSR	0.486	0.326
Method × MC × PO × PSR	0.408	0.458
Method × PO × PSR	0.304	0.320
Method × MC × PO	0.191	0.086
Method × MC × PSR × NO	0.113	0.101
Method × PSR × NO	0.085	0.104
Method × MC × PO × PSR × NO	0.026	0.023
Method × PO × PSR × NO	0.022	0.019
Method × MC × NO	0.014	0.022
Method × NO	0.009	0.087
Method × PO × NO	0.005	0.004
Method × MC × PO × NO	0.002	0.001

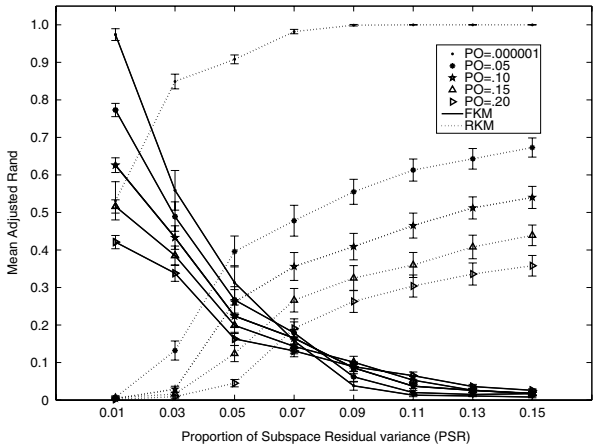

Fig. 5. Means and associated 95% Confidence Intervals of Adjusted Rand Indices as a function of Proportion of Subspace Residual variance (PSR) and Proportion of Overlap (PO).

To identify factors with a strong influence on the subspace recovery, a full-factorial RMANOVA was performed. As can be seen in Table 1, relatively strong effects were found for Method and for almost all interactions of Method with PO, PSR, and MC. Note that all effects involving the Number of Objects appear to be weak (partial $\eta^2 < 0.20$). A careful examination of the strong effects revealed the following: For RKM, analyses of data for which ($Q = 1, C = 2$) appeared to generally yield better subspace recovery than analyses of data for which ($Q = 2, C = 4$), whereas the reverse holds for FKM. The interaction between PO, PSR and Method is summarized in Fig. 4. In Fig. 4, it can be seen that for RKM at a given level of relative size of subspace residual variance (PSR), the quality of subspace recovery decreases with increasing overlap between clusters. For FKM, the opposite effect is found, i.e., at a given level of PSR, the quality increases with increasing overlap. The latter appears counterintuitive. Apparently, the quality of subspace recovery by FKM increases with increasing relative sizes of complement residuals compared to subspace residuals, as we conjectured, but also with decreasing relative sizes of the centroid scores compared to the sizes of the complement residuals.

A Phi coefficient lower than, say, 0.60 already indicates a rather bad recovery of the loadings. Therefore, FKM performs generally reasonably in the conditions with PSR = 0.01 and 0.03, whereas the quality of the RKM solution is acceptable for PSR levels higher than 0.11.

As indicated by the RMANOVA the number of objects in the range examined (50 to 200) affects the subspace recovery relatively little (partial $\eta^2 < 0.11$). Inspection of the results revealed that performance generally improves with increasing number of objects, whereas FKM benefits most from increased sample size, especially in conditions with relatively good performance.

Recovery of the cluster memberships

To offer insight into the cluster recovery performance of RKM and FKM, Fig. 5 presents the means (and associated 95% CIs) of the Adjusted Rand Indices as a function of Proportion of Subspace Residual variance (PSR) and Proportion of Overlap.

It can be seen in Fig. 5 that the performance of FKM in recovery of the clustering structure increases with increasing relative sizes of complement residuals (i.e., decreasing PSR), whereas this recovery decreases for RKM, as had been

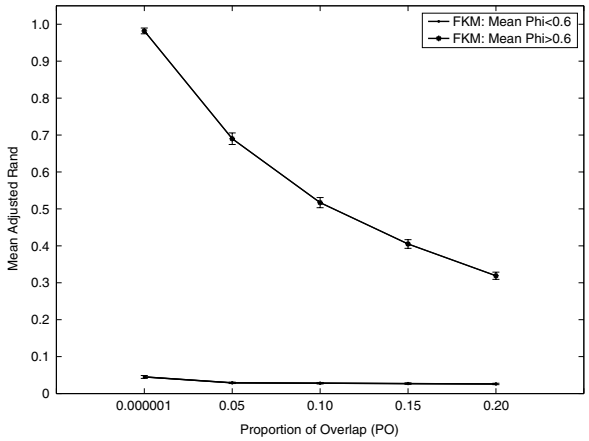

Fig. 6. FKM: Means and associated 95% Confidence Intervals of Adjusted Rand Indices as a function of Proportion of Overlap (PO), and Subspace Recovery Reasonableness.

hypothesized. Furthermore, as conjectured, for FKM as well as RKM, cluster recovery generally deteriorates with increasing amount of overlap.

A full factorial RMANOVA was performed to reveal the extent to which the manipulated factors influence membership recovery. As can be seen in Table 1, also membership recovery depends mainly on the method used and on the interactions of the method with PO, PSR, and MC. Inspection of the results indicated that the model complexity affected the membership recovery of FKM and RKM in a similar vein as it did for the subspace recovery. The interaction effect of PO, PSR and Method on membership recovery is summarized in Fig. 5. As stated above, the membership recovery generally tends to decrease with increasing overlap between clusters, for each of level of PSR, which is understandable from a theoretical point of view. For FKM, this pattern differs from the one found for the subspace recovery, which tended to ameliorate with increasing overlap. In fact, we can make a distinction between cases for which the correct subspace has been identified, and those for which it has not. In the former case, the membership recovery is easier when clusters overlap less—if the correct subspace has not been identified, the membership recovery is arbitrary anyway. This hypothesis is confirmed by further considering those cases for which FKM has a reasonable recovery of the subspace, and those for which it has not, that is for which Mean Phi $>$ 0.60 and for which this is $<$0.60. As can be seen in Fig. 6, in cases for which Mean Phi $>$ 0.60, increasing PO strongly affects the recovery of the memberships, whereas in cases with Mean Phi $<$ 0.60 the recovery of membership is around the chance level.

4.5. Implications of the simulation results for empirical practice

In this simulation study, we studied how well RKM and FKM succeed in recovering the correct subspace and the correct cluster membership. The results of this study confirmed our expectation that the recovery of the cluster memberships by FKM and RKM deteriorates with increasing overlap between clusters. Furthermore, confirmation was found for our conjecture that the quality of subspace recovery increases for FKM with decreasing relative sizes of subspace residuals in comparison to the complement residuals, and that the reverse holds for RKM. An important question is what the implications of those findings are for the empirical use of FKM and RKM.

The key difference in performance between FKM and RKM appears to be in the subspace recovery. After all, once the proper subspace has been identified, FKM and RKM perform equally well, and are affected in a similar vein by increasing overlap between clusters. The decrease in performance with increasing overlap is not surprising, and will generally occur, irrespective of the clustering method considered.

With respect to subspace recovery, FKM and RKM appear to complement each other in finding the proper subspace: FKM may provide a proper solution when RKM fails, and vice versa. To judge which method is most likely to yield a proper recovery for an empirical data set, one has to assess the proportion of subspace residual variance. In doing so, it is important to realize that masking variables are an important source of complement residual variance, because they fully lie in the complement subspace. For variables that do reflect the clustering structure, it is hard to predict whether the residuals are in the clustering subspace or in its complement.

In our simulation study RKM appeared to perform reasonably well in subspace recovery when the proportion of subspace residual variance (PSR) was larger than 0.11. Although very hard to substantiate, we guess that PSRs of 0.11 or lower will occur relatively little in empirical practice, implying that RKM would have a wide applicability. Especially when the majority of the variables reflect the clustering structure, the PSR will be large enough to expect a proper performance of RKM.

On the other hand, FKM performed reasonably well if the PSR was smaller than 0.03, implying that more than 97% of the variance is in the subspace complement to the centroid subspace. This seems to be a relatively large amount, especially because variables are commonly standardized to have unit variances, resulting in a downweighting of masking variables

Table 2
Values of fit(**X**), fit(**XA**) and Adjusted Rand Indices for each FKM and RKM solution computed for the archetypal psychiatric patient data.

Q	RKM			FKM		
	fit(**X**)	fit(**XA**)	Adjusted Rand	fit(**X**)	fit(**XA**)	Adjusted Rand
1	0.411	0.968	0.446	0.002	0.983	−0.005
2	0.529	0.865	0.691	0.007	0.919	0.103
3	0.639	0.846	0.876	0.012	0.816	0.121
4	0.639	0.788	0.876	0.015	0.745	−0.005
5	0.639	0.834	0.876	0.014	0.598	−0.011
6	0.639	0.796	0.876	0.018	0.547	0.002
7	0.639	0.778	0.876	0.032	0.595	0.162
8	0.639	0.763	0.876	0.028	0.497	0.103
9	0.639	0.747	0.876	0.032	0.455	0.066
10	0.639	0.727	0.876	0.050	0.503	0.258
11	0.639	0.706	0.876	0.079	0.548	0.349
12	0.639	0.684	0.876	0.055	0.390	0.152
13	0.639	0.676	0.876	0.091	0.450	0.101
14	0.639	0.663	0.876	0.415	0.741	0.380
15	0.639	0.650	0.876	0.459	0.704	0.511
16	0.639	0.646	0.876	0.554	0.687	0.773
17	0.639	0.639	0.876	0.639	0.639	0.876

with large variances. When in empirical practice no educated guess whatsoever can be made on the relative size of PSR, it appears wise to consider both RKM and FKM.

5. Empirical example: The archetypal psychiatric patient data

To illustrate the use of FKM and RKM in empirical practice, we consider data from a study on archetypal psychiatric patients (Mezzich and Solomon, 1980). In this study each of 11 psychiatrists was invited to describe a typical patient for each of four diagnostic categories: manic-depressive depressed (MDD), manic-depressive manic (MDM), simple schizophrenic (SS) and paranoid schizophrenic (PS). These diagnostic categories are part of the DSM-II nomenclature of mental disorders (American Psychiatric Association, 1968). The 11 psychiatrists characterized each of the four archetypal patients by rating the 17 symptoms from the Brief Psychiatric Rating Scale (BPRS) on a seven point scale (ranging from 1 = not present to 7 = extremely severe), resulting in a 44×17 data matrix **X**. The following three questions arise: Can the four archetypal patients be recovered from the data? Which variables are important in defining the four diagnostic categories under study? Finally, can the 17 symptoms be reduced to a few meaningful dimensions?

The standardized version of the 44×17 data matrix **X** was analyzed with both RKM and FKM with 4 clusters and the number of components (Q) ranging from 1 to 17, where for each analysis 1000 random starts were used. For each solution, we computed the percentages of variance explained by the estimated model of both the observed data (**X**), and of the projected version of the observed data (**XA**), indicated by fit(**X**) and fit(**XA**), respectively. Also, the quality of recovery of the four archetypal patients was assessed by considering the Adjusted Rand index between the estimated and expected membership matrices.

5.1. Selection of FKM and RKM solutions

As can be seen in Table 2, for RKM the fit of the observed data (fit(**X**)) is non-decreasing with increasing number of components, as it should be. To select the optimal number of components for RKM on the basis of fit only (i.e., without considering the usually unknown Adjusted Rand Index values), one could apply the scree criterion to a plot of the fit(**X**)-values against the Q-values (Cattell, 1966), so as to identify the model with an optimal balance of fit and complexity. This scree criterion clearly suggests the selection of a RKM solution with $Q = 3$ components. The Adjusted Rand index values of the RKM solutions (in Table 2) further reveal that adding more components would not improve the clustering. Hence, the RKM solution with $Q = 3$ components is the most parsimonious solution with a reasonable fit and a reasonable clustering recovery, and is therefore considered to be the optimal RKM solution.

For selecting among FKM solutions, the scree criterion cannot be used, since the fit of FKM of the projected data often decreases with increasing numbers of components; this is due to the fact that the projected data **XA** commonly varies with differing numbers of components. For FKM, the fit of the observed data, fit(**X**), is of rather limited informative value. Indeed, low fit(**X**)-values may either indicate that FKM is highly successful in ignoring the masking variables when clustering the objects, or that FKM fails to model the clustering structure present in the data. To date, no satisfactory approach appears to exist to select the number of components in FKM analysis.

For the archetypal psychiatric patient data, the Adjusted Rand Index values of the various FKM solutions indicate that FKM poorly recovers the clustering structure; only with $Q = 16$ or 17 components does the clustering structure appear to be recovered to a reasonable extent. As these solutions are much more complex than the optimal RKM solution, only the optimal RKM solution will be considered in the next paragraphs.

Table 3
Rotated centroid scores of the $C = 4$ clusters on the $Q = 3$ components of the RKM solution; MDD is manic-depressive depressed, MDM is manic-depressive manic, SS is simple schizophrenic and PS is paranoid schizophrenic.

	I (MDM)	II (PS)	III (SS)
1 (MDM)	**0.50**	0.04	−0.14
2 (PS)	0.02	**0.44**	−0.18
3 (SS)	−0.08	−0.09	**0.42**
4 (MDD)	**−0.37**	**−0.39**	−0.05

Table 4
Rotated loadings of 17 variables on the $Q = 3$ components of the RKM solution; MDD is manic-depressive depressed, MDM is manic-depressive manic, SS is simple schizophrenic and PS is paranoid schizophrenic. Loadings in absolute value > 0.20 are indicated in bold face.

	I (MDM)	II (PS)	III (SS)
Excitement	**0.33**	0.05	**−0.25**
Grandiosity	**0.26**	0.18	**−0.24**
Somatic concern	**−0.40**	0.00	−0.18
Anxiety	**−0.35**	0.08	−0.38
Emotional withdrawal	**−0.38**	0.10	**0.25**
Motor retardation	**−0.26**	**−0.28**	−0.09
Depressive mood	**−0.28**	**−0.27**	**−0.23**
Guilt feelings	−0.26	**−0.28**	**−0.26**
Mannerisms and posturing	−0.14	**0.32**	**0.28**
Hostility	0.08	**0.28**	**−0.21**
Suspiciousness	−0.19	**0.39**	−0.17
Hallucinatory behavior	−0.17	**0.40**	−0.04
Uncooperativeness	0.01	**0.28**	−0.21
Unusual thought content	−0.18	**0.31**	−0.10
Conceptual disorganization	0.02	**0.22**	0.15
Blunted affect	−0.17	0.06	**0.46**
Tension	0.19	0.06	**−0.27**

5.2. Interpretation of the selected RKM model

To facilitate the interpretation of the selected RKM solution, we orthogonally rotated the centroid matrix **F** towards simplicity using the normalized Varimax criterion (Kaiser, 1958), and compensated for this rotation in the loading matrix **A**. In this particular case, we choose to rotate the centroid matrix rather than the loading matrix, because then each cluster will be characterized by a linear combination of as few components as possible. The rotated centroid scores are presented in Table 3.

As can be seen in Table 3, the centroid scores show a simple pattern: Cluster 1 is mainly related to Component I, Cluster 2 to Component II, and Cluster 3 to Component III; Cluster 4 is mainly an about equally weighted combination of Components I and II.

To interpret the Clusters and Components further, we inspect the rotated loading matrix, which is presented in Table 4. Inspection of the loadings reveals that Component 1 shows high positive loadings on symptoms that are present with typical MDM patients and low loadings on symptoms that are absent in those patients; therefore, Component 1 is labeled as MDM. In a similar vein, Components II and III can be labeled as PS and SS, respectively. Because all variables load to a reasonable extent on at least one component, all variables are relevant in characterising the typical patients. This is not surprising, because the BPRS, from which the variables stem, is designed for diagnosing psychiatric patients.

Having labeled the Components, we can assign labels to the Clusters as well, where Clusters 1 to 4 can be labeled as MDM, PS, SS and MDD, respectively. The, thus performed labeling, on the basis of the component contents, coincides with the expected cluster labeling (i.e., based on the archetypal patients rated). Because the MDM, PS and SS clusters are largely related to a single component, their symptom pattern is reflected in the component concerned. The MDD cluster is an approximately equally negatively weighted combination of MDM and PS symptoms; the negative signs imply that symptoms that are present in either MDM or PS are largely absent in MDD, and vice versa.

To obtain further insight into the clustering provided by RKM, we plotted the observed scores projected into the estimated subspace (i.e., **XA**), and the centroids of the four clusters in Fig. 7. The variability in scorings provided by the 11 psychiatrists appears to be smallest for PS. The two misclassified descriptions of an archetypal patient stemmed from two different psychiatrists. It could be interesting to relate the misclassifications to characteristics of the psychiatrist, such as years of experience, but such background information is lacking.

From these results, we conclude that RKM recovers the four archetypal patients well from the descriptions. Furthermore, all symptoms appeared to be relevant in defining the archetypal patients. Insight was obtained into the relationships between symptoms in the four archetypal patients. Archetypal MDM, SS and PS patients appear to have a specific pattern of symptoms. In contrast, MDD patients typically show symptoms that are absent in MDM and PS, and do not have symptoms present in MDM and PS.

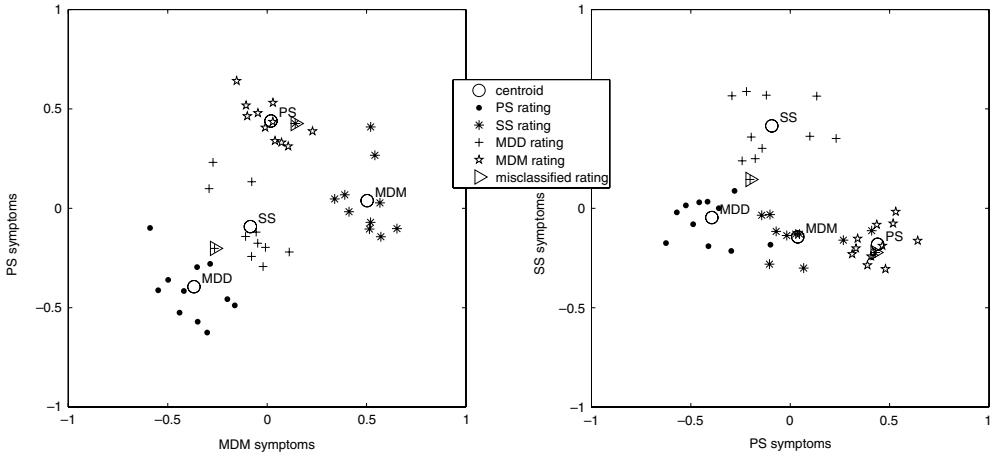

Fig. 7. Observed scores projected to the RKM subspace, for (a) the MDM versus PS dimension, and (b) the PS versus SS dimension.

6. Discussion and conclusion

This paper aimed at offering a comprehensive comparison between FKM and RKM. This knowledge is of key importance to understand the properties of FKM and RKM and to judge when either method is a proper choice to gain insight into the clustering structure of an empirical data set.

In empirical practice, the central issue to resolve boils down to a model selection problem. That is, once a choice has been made for a K-means subspace clustering approach, one has to choose between RKM and FKM, and select the numbers of components and clusters. In choosing between RKM and FKM, an educated guess is necessary on the relative size of the residuals in the clustering subspace. If such a guess is impossible to make, it appears wise to consider both RKM and FKM. However, although hard to substantiate, we conjecture that RKM generally will show a good performance when the majority of the variables reflect the clustering structure and/or the variables are standardized before analysis.

Although the final model selection should be based on substantive considerations and interpretability of the solution, a model selection heuristic can be a helpful tool to identify one or more interesting models. To this end, for RKM, a heuristic used for selecting the number of clusters such as the popular Calinski–Harabasz index (1974) or the Silhouette index (Kaufman and Rousseeuw, 1990) could be applied; alternatively, a heuristic applied in component analysis, as in the numerical convex hull approach (Ceulemans and Kiers, 2006), could be used. The latter proved a better model selection heuristic in the context of multi-mode partitioning than the Calinski–Harabasz and Silhouette indices (Schepers et al., 2008). For FKM, such heuristics cannot be used, as the fit may decrease with increasing model complexity. Further research is needed to examine the usefulness of the heuristics mentioned for choosing among RKM solutions, and to define useful FKM heuristics.

Appendix A

In Section 3.1, it was stated that the RKM loss function is zero if and only if $\hat{\mathbf{A}}$ is $\mathbf{AT}$ for some orthonormal matrix $\mathbf{T}$, $\hat{\mathbf{U}}$ and $\mathbf{U}$ differ at most by a permutation, i.e., $\hat{\mathbf{U}} = \mathbf{U\Pi}$, for some permutation matrix $\mathbf{\Pi}$, and $\hat{\mathbf{F}} = \mathbf{\Pi}'\mathbf{FT}$, provided that rank$(\mathbf{U}) = C$ and $\mathbf{F}$ has rank $Q(Q \leq C)$ and does not have any equal rows, and rank$(\mathbf{A}) = Q, Q \leq \min(C, J)$. This can be proven as follows. From $\mathbf{UFA}' = \hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}'$ we have $\mathbf{UF} = \hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}$. Furthermore, because $\mathbf{F}$ does not have any equal or zero rows, $\mathbf{UF}$ is known to have exactly C different rows, or in other words, contains C sets of equal rows. The estimated membership matrix $\hat{\mathbf{U}}$ has at most C different rows, and hence at most C sets of equal rows. Now, it will be explained that the sets of equal rows must be the same in $\mathbf{UF}$ and $\hat{\mathbf{U}}$. This is because, on the one hand, if in $\hat{\mathbf{U}}$ two rows were equal while the corresponding rows were not equal in $\mathbf{UF}$, the equality $\mathbf{UF} = \hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}$ could never hold. Hence rows that are equal in $\hat{\mathbf{U}}$ must also be equal in $\mathbf{UF}$. On the other hand, if in $\hat{\mathbf{U}}$ two rows were different while the corresponding rows in $\mathbf{UF}$ were equal, this would imply that two different *sets* of rows in $\hat{\mathbf{U}}$ must correspond to a single set of equal rows in $\mathbf{UF}$. Since there are at most C different sets of rows in $\hat{\mathbf{U}}$, this would imply that these sets jointly correspond to at most $C - 1$ different sets of rows in $\mathbf{UF}$, which, however conflicts with the fact that $\mathbf{UF}$ is known to have exactly C different rows. Therefore, the sets of equal rows must be the same in $\mathbf{UF}$ and $\hat{\mathbf{U}}$.

From the notion that the sets of equal rows must be the same in $\mathbf{UF}$ and $\hat{\mathbf{U}}$, it follows that $\mathbf{U}$ and $\hat{\mathbf{U}}$ can differ at most by a permutation of the columns. As a consequence, $\hat{\mathbf{U}} = \mathbf{U\Pi}$, for some permutation matrix $\mathbf{\Pi}$, and $\hat{\mathbf{F}}\hat{\mathbf{A}}' = \mathbf{\Pi}'\mathbf{FA}'$. Because $\mathbf{F}$ has rank Q, the latter equality implies that $\mathbf{A}$ and $\hat{\mathbf{A}}$ must span the same column space, and since they are both columnwise orthonormal, they can differ at most by an orthonormal transformation matrix $\mathbf{T}$. With these results it follows from $\hat{\mathbf{F}} = \mathbf{\Pi}'\mathbf{FT}$ that $\mathbf{UFA}' = \hat{\mathbf{U}}\hat{\mathbf{F}}\hat{\mathbf{A}}'$.

Appendix B

In Section 3.2, it was stated that when $\text{rank}(\mathbf{U}) = C$, $\text{rank}([\mathbf{UF}|E^{\perp}]) = (Q+J)$, where $[\mathbf{UF}|\mathbf{E}^{\perp}]$ is a block matrix consisting of matrices $\mathbf{UF}$ and $\mathbf{E}^{\perp}$ positioned next to each other, $\text{rank}(\mathbf{A}) = Q$, $Q \leq J$, $\mathbf{F}$ does not have any equal rows, and $\mathbf{E}^{\perp}$ has no clustering structure whatsoever, that $\hat{\mathbf{A}}$ is $\mathbf{AT}$, $\hat{\mathbf{U}}$ and $\mathbf{U}$ differ at most by a permutation, i.e., $\hat{\mathbf{U}} = \mathbf{U\Pi}$, for some permutation matrix $\mathbf{\Pi}$, and $\hat{\mathbf{F}} = \mathbf{\Pi}'\mathbf{FT}$. This can be proven as follows.

In the case of perfect fit of the FKM model, we have $\mathbf{X}\hat{\mathbf{A}} = (\mathbf{UFA}' + \mathbf{E}^{\perp}\mathbf{A}^{\perp\prime})\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\mathbf{F}}$. From this, it follows that

$$[\mathbf{UF}|\mathbf{E}^{\perp}][\mathbf{A}|\mathbf{A}^{\perp}]'\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\mathbf{F}}, \tag{13}$$

hence $[\mathbf{UF}|\mathbf{E}^{\perp}]$ and $\hat{\mathbf{U}}\hat{\mathbf{F}}$ span the same column spaces. Because $\mathbf{U}$ and $\hat{\mathbf{U}}$ both have clustering structure and $\mathbf{E}^{\perp}$ has not, it follows that the linear combinations that constitute the columns of $\hat{\mathbf{U}}\hat{\mathbf{F}}$ from those of $[\mathbf{UF}|\mathbf{E}^{\perp}]$ must be based only on the former part of the matrix, and hence use zero weights in the latter part. In other words, this implies that within $([\mathbf{A}|\mathbf{A}^{\perp}]'\hat{\mathbf{A}})$ the part $(\mathbf{A}^{\perp\prime}\hat{\mathbf{A}})$ equals $\mathbf{0}$. It follows that $\hat{\mathbf{A}}$ must fully reside in the column space of $\mathbf{A}$, and, because they are both columnwise orthonormal, they can differ at most by an orthonormal transformation matrix $\mathbf{T}$. Substituting $\hat{\mathbf{A}} = \mathbf{AT}$ in (7) we get $\mathbf{UFA}'\mathbf{AT} = \mathbf{UFT} = \hat{\mathbf{U}}\hat{\mathbf{F}}$, from which analogously to the reasoning in Appendix A it follows that $\hat{\mathbf{U}} = \mathbf{U\Pi}$, for some permutation matrix $\mathbf{\Pi}$, and $\hat{\mathbf{F}} = \mathbf{\Pi}'\mathbf{FT}$.

References

American Psychiatric Association, 1968. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association, Washington, DC.

Arabie, P., Hubert, L., 1994. Cluster analysis in marketing research. In: Bagozzi, R.P. (Ed.), Handbook of Marketing Research. Blackwell, Oxford.

Bock, H.H., 1987. On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In: Bozdogan, H., Gupta, A.K. (Eds.), Multivariate Statistical Modeling and Data Analysis. D. Reidel Publishing Company, Dordrecht, pp. 17–34.

Calinski, R.B., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics 3, 1–27.

Cattell, R.B., 1966. The scree test for the number of factors. Multivariate Behavioral Research 1, 245–276.

Ceulemans, E., Kiers, H.A.L., 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, 133–150.

Cliff, N., 1966. Orthogonal rotation to congruence. Psychometrika 31, 33–42.

De Soete, G., Carroll, J.D., 1994. K-means clustering in a low-dimensional Euclidean space. In: Diday, E., et al. (Eds.), New Approaches in Classification and Data Analysis. Springer, Heidelberg, pp. 212–219.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.

Kaiser, H.F., 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 187–200.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: LeCarn, L.M., Neyman, J. (Eds.), 5th Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 281–296.

Mezzich, J.E., Solomon, H., 1980. Taxonomy and Behavioral Science. Academic Press, London.

Milligan, G.W., 1996. Clustering validation: Results and implications for applied analysis. In: Arabie, P., Hubert, L.J., De Soete, G. (Eds.), Clustering and Classification. World Scientific Publishing, River Edge, pp. 341–375.

Milligan, G.W., Cooper, M.C., 1988. A study of standardization of variables in cluster analysis. Journal of Classification 5, 181–204.

Schepers, J., Ceulemans, E., Van Mechelen, I., 2008. Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria. Journal of Classification 25, 67–85.

Steinley, D., Brusco, M.J., 2008. Selection of variables in cluster analysis: An empirical comparison of eight procedures. Psychometrika 73, 125–144.

Steinley, D., Henson, R., 2005. OCLUS: An analytic method for generating clusters with known overlap. Journal of Classification 22, 221–250.

Tucker, L.R., 1951. A method for synthesis of factor analysis studies, Personnel Research Section Report No. 984. Department of the Army, Washington, DC.

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