Journal of Classification 29:297-320 (2012)
DOI: 10.1007/s00357-012-9112-5

Lowdimensional Additive Overlapping Clustering

Dirk Depril
suAzio Consulting, Belgium

Iven Van Mechelen
KU Leuven, Belgium

Tom F. Wilderjans
KU Leuven, Belgium

Abstract: To reveal the structure underlying two-way two-mode object by variable data, Mirkin (1987) has proposed an additive overlapping clustering model. This model implies an overlapping clustering of the objects and a reconstruction of the data, with the reconstructed variable profile of an object being a summation of the variable profiles of the clusters it belongs to. Grasping the additive (overlapping) clustering structure of object by variable data may, however, be seriously hampered in case the data include a very large number of variables. To deal with this problem, we propose a new model that simultaneously clusters the objects in overlapping clusters and reduces the variable space; as such, the model implies that the cluster profiles and, hence, the reconstructed data profiles are constrained to lie in a lowdimensional space. An alternating least squares (ALS) algorithm to fit the new model to a given data set will be presented, along with a simulation study and an illustrative example that makes use of empirical data.

Keywords: Additive overlapping clustering; Dimensional reduction; Alternating least squares algorithm; Two-way two-mode data; Object by variable data.

The research in this paper was partially supported by the Research Fund of KU Leuven (PDM-kort project 3H100377, dr. Tom F. Wilderjans; GOA $2005/04$, Prof. dr. Iven Van Mechelen), by the Belgian Science Policy (IAP $P6/03$, Prof. dr. Iven Van Mechelen), and by the Fund of Scientific Research (FWO)-Flanders (project G.0546.09, Prof. dr. I. Van Mechelen). The simulation study was conducted using high performance computational resources provided by the KU Leuven (http://ludit.kuleuven.be/hpc). Requests for reprints should be sent to Tom F. Wilderjans. The authors are obliged to Prof. dr. Peter Kuppens for kindly providing the data of Section 5 and to the anonymous reviewers for most helpful remarks on previous versions of this paper.

Corresponding Author's Address: Tom F. Wilderjans, Faculty of Psychology and Educational Sciences, KU Leuven, Andreas Vesaliusstraat 2, Box 3762, B-3000 Leuven, Belgium, Tel: +32. 16. 32.61.23, Fax: +32. 16. 32.62.00, e-mail: tom.wilderjans@ppw.kuleuven.be

Published online 17 August 2012

1. Introduction

A rectangular $I \times J$ object by variable data array $\mathbf{X}$ is often encountered in statistical practice. Given such data, one may wish to know its structure and the underlying mechanism that generated it. In several cases, hypothesized structures may be based on a number of groupings of the objects under study, each grouping being associated with some substantive aspect of the objects. For instance, in case of a patient by symptom data set, a grouping may refer to a certain syndrome; in case of a consumer by brand data set, a grouping may refer to a certain consumer need. To discover such groupings of objects on the basis of a data set, one may rely on some form of cluster analysis.

In practice, the type of cluster structure that most often is looked for is a partitioning, which implies that each object belongs to a single group only. In some cases, however, a partitioning may not correspond to the true structure underlying the data at hand because the latent object aspects may overlap. As examples, patients sometimes may suffer from more than a single disease, and a single consumer may have multiple consumer needs.

In such cases, models that imply an overlapping clustering of the objects should be used instead of a partitioning. For the special case of two-way two-mode (i.e., object by variable) data, an additive overlapping clustering model has been proposed by Mirkin (1987). Along with the overlapping object clusters, Mirkin's model includes a set of vectors, called cluster profiles, which comprise the characteristic values on each variable for the clusters in question. The model variable values of an object then equals the sum of the profiles of the clusters the object in question belongs to. As such, the model of Mirkin can be conceived as a variant for two-way two-mode (i.e., object by variable) data of the well-known ADCLUS model of Shepard and Arabie (1979), which has been proposed to disclose overlapping object clusters from two-way one-mode (i.e., object by object) similarity data[1]. Note that Mirkin's model also subsumes the widely used K-means model as a special case. An effective algorithm to fit Mirkin's additive overlapping clustering model to an empirical two-way two-mode data set has been presented by Depril, Van Mechelen, and Mirkin (2008); software for fitting Mirkin's model, using the algorithm(s) presented in Depril et al. (2008), has been proposed by Wilderjans, Ceulemans, Van Mechelen, and Depril (2011).

1. Some authors refer to Mirkin's model as a one-mode additive overlapping clustering model (for two-way two-mode data), with 'one-mode' here indicating that only for the elements of one data mode (i.e., the objects) an overlapping clustering is obtained.

The profiles of Mirkin's additive overlapping clustering model are the key to interpret the clusters as included in the model. However, the more variables are involved in the data, the more difficult this interpretation becomes. This can become especially troublesome in case of highdimensional data. As a way out in such cases, one could try to reduce the set of variables to a much smaller set by means of some dimension reduction technique. This idea gives rise to the problem of looking simultaneously for a clustering and a variable reduction that capture in an (jointly) optimal way the structure behind the data at hand.

The problem of looking for an optimal clustering along with an optimal variable reduction has already been addressed in the case of partitioning methods (Tryon and Baily 1970; Everitt 1977). One possible solution to this problem that has been proposed in that case is to sequentially apply a clustering and a dimension reduction one after the other (Tryon and Baily 1970; Everitt 1977). In particular, one could first cluster the objects and subsequently perform a dimension reduction technique on the resulting centroids; as an alternative, one could first reduce the variables, by taking the first few principal components, and subsequently perform a clustering on the object scores on these components. Sequential methods, however, have been subjected to several kinds of criticism. On the one hand, a number of variables in $\mathbf{X}$ may not be informative about the clustering as present in the data, or, in the worst case, may even mask the underlying clustering; as a consequence, performing first a clustering in such a case may yield an incorrect clustering (Vichi and Kiers 2001). On the other hand, applying first a PCA is strongly disadviced by Arabie and Hubert (1994) and Chang (1983), since the first principal components may not be informative about the cluster structure in the data; this is especially the case when the clustering is present in the directions with the smallest eigenvalues (Vichi and Kiers 2001). To overcome the problems implied by sequential approaches, in a number of recent papers methods have been proposed that look for a partitioning and a dimension reduction simultaneously, see, for two-way models, Bock (1987), De Soete and Carroll (1994), and Vichi and Kiers (2001); for three-way models, see Carroll and Chaturvedi (1995), Rocci and Vichi (2005), and Vichi, Rocci, and Kiers (2007).

For overlapping clusterings, however, at this moment no models or methods are available yet to perform a clustering simultaneously with a dimensional reduction. Therefore, in the present paper we will propose a novel model and algorithm that performs an overlapping object clustering along with a variable reduction. Taking into account the warnings against sequential approaches in the case of partitioning methods, the novel method will aim at finding the clustering and variable reduction simultaneously. Our new model will be a constrained version of the additive overlapping cluster-

ing model of Mirkin (1987). In particular, in the new model the cluster profiles of Mirkin's model will be restricted to lie in a lowdimensional subspace of the data space.

The remainder of this paper is structured as follows. First, in Section 2, we will propose the lowdimensional additive overlapping clustering model. Next, in Section 3, the estimation of the model will be explained. In the subsequent sections a simulation study (Section 4) and an application (Section 5) will be presented. We will conclude with a discussion in Section 6.

2. The Lowdimensional Additive Overlapping Clustering Model

In this section, we will develop a model that establishes a clustering of the objects and a dimensional reduction of the variables simultaneously. The new model will be a constrained version of the additive overlapping clustering model proposed by Mirkin (1987). First, we will recapitulate this model in the next paragraph.

The additive overlapping clustering model proposed by Mirkin (1987) aims at clustering the objects of a two-way two-mode $I \times J$ object by variable data matrix $\mathbf{X}$. The data $\mathbf{X}$ are approximated by an $I \times J$ model matrix $\mathbf{M}$ that, in turn, can be decomposed into a binary $I \times K$ matrix $\mathbf{A}$ and a real-valued $K \times J$ matrix $\mathbf{P}$, with

$$\mathbf{M} = \mathbf{AP}, \tag{1}$$

that is,

$$m_{ij} = \sum_{k=1}^{K} a_{ik} p_{kj}, \tag{2}$$

with $K \leq I$ being the smallest number for which such a decomposition is possible. The matrix $\mathbf{A}$ is called the *cluster membership matrix*, the columns of which define K clusters, with entry a_{ik} denoting whether object i belongs to cluster k ($a_{ik} = 1$) or not ($a_{ik} = 0$). Apart from the restriction of binary entries, no further constraints are put on the entries of $\mathbf{A}$; this implies that the clustering as defined by $\mathbf{A}$ can be an overlapping one. The matrix $\mathbf{P}$ is called the *profile matrix*, the rows of which constitute the *variable profiles* of the clusters, with entry p_{kj} denoting the value of variable j for cluster k (the term profile is used here instead of the more classical term centroid, to point out that the rows of $\mathbf{P}$ do not have to coincide with the mean vectors of the objects in the corresponding clusters). Equation (2) then implies that the value of object i for variable j is the sum of the values for the j-th variable as included in the profiles of the clusters object i belongs to.

To achieve a dimensional reduction of the variables simultaneously with the clustering of the objects, we will constrain the profiles of Mirkin's

additive overlapping clustering model to lie in a lowdimensional space, that is equation (1) holds with the restriction that

$$S \leq \min(K, J), \tag{3}$$

with S denoting the rank of $\mathbf{P}$.

Rank constraint (3) implies that one can decompose the profile matrix $\mathbf{P}$ as

$$\mathbf{P} = \mathbf{CB}'. \tag{4}$$

The $J \times S$ matrix $\mathbf{B}$ contains in its columns base vectors of the lowdimensional row space, and the $K \times S$ matrix $\mathbf{C}$ contains the scores of the profiles on these base vectors. The entries of $\mathbf{B}$ can be used to interpret the base vectors in terms of the J original variables, and the entries of $\mathbf{C}$ then can be used to interpret the clusters. Note that decomposition (4) is only unique upon a rotation of the base vectors, as for every rotation matrix $\mathbf{U}$ it holds that $\mathbf{P} = \mathbf{CUU}'\mathbf{B}'$.

One may note that the rank constraint on $\mathbf{P}$ also implies that $\mathrm{rank}(\mathbf{M}) = S$. In general, it holds for any overlapping clustering model (1) that $\mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\mathbf{M})$, provided that K is minimal. To show this, we first prove that $\mathbf{A}$ is always of full rank. Indeed, if $\mathbf{A}$ would be rank deficient, a non-null vector $\mathbf{t}$ exists such that $\mathbf{0} = \mathbf{At}$. Given that $\mathbf{t}$ is non-null, we can assume, without loss of generality, that the first entry of $\mathbf{t}$ differs from zero. Consider then a $K \times K$ identity matrix and replace its first column by $\mathbf{t}$. It is easy to see that the resulting $K \times K$ matrix $\mathbf{T}$ is invertible. It therefore holds that $\mathbf{M} = \mathbf{AP} = \mathbf{ATT}^{-1}\mathbf{P} = \tilde{\mathbf{A}}\tilde{\mathbf{P}}$ with $\tilde{\mathbf{A}} = \mathbf{AT}$ and $\tilde{\mathbf{P}} = \mathbf{T}^{-1}\mathbf{P}$. Since $\mathbf{At} = \mathbf{0}$, the first column of matrix $\tilde{\mathbf{A}}$ is a zero column, which implies that the model $\mathbf{M}$ can be represented with $K - 1$ clusters. Therefore, the constraint of K being minimal (as associated with model (1)) is violated. It thus holds that $\mathbf{A}$ is of full rank, provided that K is minimal.

We now turn to the actual proof that for any overlapping clustering model (1) it holds that $\mathrm{rank}(\mathbf{M}) = \mathrm{rank}(\mathbf{P})$. Firstly, $\mathrm{rank}(\mathbf{M}) \leq \mathrm{rank}(\mathbf{P})$, since $\mathrm{rank}(\mathbf{M}) = \mathrm{rank}(\mathbf{AP}) \leq \min(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{P}))$. Secondly, from the fact that $\mathbf{A}$ is of full rank, it follows that $\mathbf{P} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{M}$. Hence, $\mathrm{rank}(\mathbf{P}) \leq \min(\mathrm{rank}((\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'), \mathrm{rank}(\mathbf{M}))$, hence $\mathrm{rank}(\mathbf{P}) \leq \mathrm{rank}(\mathbf{M})$.

3. Estimation

3.1 Loss Function

Given an empirical $I \times J$ data matrix $\mathbf{X}$, it is always possible to exactly represent it in terms of a lowdimensional additive overlapping clustering model $\mathbf{M}$. In practice, however, one may wish to capture the structure behind the data, rather than to go for a perfect data reconstruction. As such,

one may look for an approximate representation of $\mathbf{X}$ for relatively small values of K and S. In particular, we allow for discrepancies between the data matrix $\mathbf{X}$ and the model matrix $\mathbf{M}$, which means that

$$\mathbf{X} = \mathbf{M} + \mathbf{E}. \tag{5}$$

A model $\mathbf{M} = \mathbf{AP}$ with prespecified values of $K < I$ and $S < \min(K, J)$ will then be estimated by minimizing the least squares loss function

$$L^2(\mathbf{A}, \mathbf{P}) = \|\mathbf{X} - \mathbf{AP}\|_F^2 = \sum_{i,j=1}^{I,J} (x_{ij} - \sum_{k=1}^{K} a_{ik} p_{kj})^2 \tag{6}$$

over $\mathbf{A}$ and $\mathbf{P}$, and subject to $\mathbf{A}$ being binary and $\mathbf{P}$ being real-valued with $\mathrm{rank}(\mathbf{P}) = S$. Remark that $\|.\|_F$ denotes the Frobenius norm of a matrix (i.e., the square root of the sum of the squared entries).

In practice, the number of clusters K and the number of dimensions of the subspace S are usually unknown. To determine these, one could fit a series of models corresponding to a range of values for K and S, and subsequently select one of them on the basis of some model selection heuristic, like the CHull method (Wilderjans, Ceulemans, and Meers 2012) or a generalization of the scree test (Cattell 1966) as implemented in Wilderjans, Ceulemans, and Kuppens (2012); alternative heuristics for similar model selection problems are described in Ceulemans, Van Mechelen, and Leenen (2003), Ceulemans and Van Mechelen (2005), Schepers, Ceulemans, and Van Mechelen (2008), and Wilderjans, Ceulemans, and Van Mechelen (in press). Also, one could decide on K and S on the basis of substantive considerations.

3.2 Algorithm

Since the membership matrix $\mathbf{A}$ is binary, there is no closed form solution for the minimization of the loss function (6). For this minimization, we will therefore propose an alternating least squares (ALS) algorithm. Starting from an initial membership matrix $\mathbf{A}_0$, this algorithm will estimate the conditionally optimal profiles $\mathbf{P}$ upon $\mathbf{A}$; subsequently it will estimate the conditionally optimal memberships $\mathbf{A}$ upon $\mathbf{P}$, and this process will be repeated until convergence.

The initial membership matrix $\mathbf{A}_0$ can be either a priori given or randomly drawn. For the latter case, we propose two different strategies. The first one is to take independent draws from a Bernoulli distribution with parameter $\pi = .5$ (i.e., randomly sampled $\mathbf{A}_0$). The second one is to calculate the conditionally optimal membership matrix that goes with a profile ma-

trix as constituted by K randomly drawn data points (i.e., semi-randomly determined $\mathbf{A}_0$).

The optimal profile matrix $\hat{\mathbf{P}}$, given a membership matrix $\mathbf{A}$, is calculated by means of Reduced Rank Regression (Anderson 1951; Stoica and Viberg 1996). In particular, it is given by

$$\hat{\mathbf{P}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{T}\mathbf{T}'\mathbf{X}, \tag{7}$$

with the columns of the matrix $\mathbf{T}$ being the first S orthonormal eigenvectors of the matrix $\mathbf{Z}\mathbf{X}\mathbf{X}'\mathbf{Z}$, with $\mathbf{Z} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ being the orthogonal projector operator (i.e., $\mathbf{A}\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{T}\mathbf{T}'\mathbf{X} = \mathbf{Z}\mathbf{T}\mathbf{T}'\mathbf{X}$).

For the estimation of a new conditionally optimal membership matrix $\hat{\mathbf{A}}$, given a profile matrix $\mathbf{P}$, a separability property (Chaturvedi and Carroll 1994) of the loss function (6) may be used. Indeed, the loss function can be written as

$$L^2(\mathbf{A},\mathbf{P}) = \sum_j \left(x_{1j} - \sum_{k=1}^{K} a_{1k}p_{kj}\right)^2 + \cdots + \sum_j \left(x_{Ij} - \sum_{k=1}^{K} a_{Ik}p_{kj}\right)^2. \tag{8}$$

Equation (8) implies that the contribution of the i-th row of the membership matrix has no influence on the contributions of the other rows. As a consequence, $\mathbf{A}$ can be estimated row-wise, with for each row 2^K binary 0/1 patterns being possible. These will be evaluated enumeratively, retaining the combination that yields the lowest loss value.

The alternating process generates conditionally optimal solutions and, as a consequence, produces a nonincreasing sequence of positive loss values, which necessarily converges. Moreover, convergence will be reached in a finite number of steps since the solution space for the minimization problem is finite; the latter is due to the fact that this space consists of 2^{IK} solutions only, namely each possible membership matrix along with its conditionally optimal profile matrix as implied by (7).

The result of the algorithm may be a local optimum of the loss function, rather than the global minimum. Moreover, it may strongly depend on the choice of the initial membership matrix $\mathbf{A}_0$. To remedy for this, we recommend to use a random multistart procedure, which consists of taking many starts and retaining the outcome with the lowest loss value (6) as the final one. In the special case of the K-means partitioning algorithm, Steinley and Brusco (2007) found that such a random multistart strategy outperformed a number of possible competitors. Results of Depril et al. (2008) further suggest that, in the case of Mirkin's additive overlapping clustering model, a multistart strategy based on a combination of random membership and random profile starts may yield the best results.

4. Simulation Study

4.1 Introduction

In the previous section, an algorithm has been proposed to estimate the best fitting lowdimensional additive overlapping clustering model for a given data set. In this section, we will present a simulation study to evaluate the performance of this algorithm. In this regard, we are interested in two aspects of algorithmic performance: *goodness-of-fit* and *goodness-of-recovery*. With regard to *goodness-of-fit*, we will examine whether the algorithm finds the global optimum of the loss function. Concerning *goodness-of-recovery*, we will investigate to what extent the algorithm succeeds in recovering the true structure underlying a given data set. Algorithmic performance will be evaluated on a global level as well as a function of data characteristics.

In the next subsections we will outline the design of the simulation study (4.2) and the specific evaluation criteria (4.3). The results will be presented in Subsection 4.4 and discussed in Subsection 4.5.

4.2 Design

To generate data sets $\mathbf{X}$ of size $I \times J$, we will independently generate matrices $\mathbf{A}$, $\mathbf{P}$ and $\mathbf{E}$. The rows of $\mathbf{A}$ are independently drawn from a multinomial distribution on all possible 2^K binary row patterns (with a probability of .05 for the zero pattern and with the probabilities for row patterns with a single 1 put equal to each other). The matrix $\mathbf{P}$ needs to satisfy $\text{rank}(\mathbf{P}) = S$ and is therefore generated as a product of a $K \times S$ matrix $\mathbf{C}$ and a $J \times S$ matrix $\mathbf{B}$: $\mathbf{P} = \mathbf{C}\mathbf{B}'$. The entries of $\mathbf{C}$ are *iid* generated from the uniform distribution on the interval $(-20, 20)$; the rows of the matrix $\mathbf{B}$ are orthonormal and are the left singular vectors of a $J \times S$ matrix of which the entries are *iid* drawn from a standard normal distribution. Should the rank of $\mathbf{P}$ be smaller than S, then a new matrix $\mathbf{C}$ is generated until the rank is equal to S. The matrix $\mathbf{E}$ is a real-valued $I \times J$ matrix containing noise terms; its rows are independently drawn from an equicorrelated J-variate normal distribution $N(\mathbf{0}, \Sigma_E)$ with equal variances. A data set $\mathbf{X}$ is then obtained as $\mathbf{X} = \mathbf{M} + \mathbf{E}$, with $\mathbf{M} = \mathbf{A}\mathbf{P}$.

The following design factors are manipulated on the level of the data generation.

- *Data Shape:* The data shape of $\mathbf{X}$ is defined as the ratio I/J and will take three different levels: 4/1, 1/1 and 1/4. The number of entries of $\mathbf{X}$ will be kept constant to 4,096, implying three different values for $I \times J$: 128×32, 64×64, and 32×128.
- *Number of clusters K:* 4, 5, 6.

- *Amount of Cluster Overlap:* This is defined as the probability of belonging to more than one cluster and it is put equal to 25%, 50%, or 75%; for this purpose, the multinomial probabilities of all row patterns in $\mathbf{A}$ that contain more than a single 1 were put equal to one another with a total equal to the percentage in question.
- *Subspace Dimension S:* 1, 2, 3.
- *Noise level ε:* This is defined as the proportion ε of the total variance in the data $\mathbf{X}$ accounted for by $\mathbf{E}$. It is controlled through the value of the variances on the diagonal of Σ_E. This proportion ε will be either 0, .10, .20, or .30.
- *Noise correlation:* The correlations of the equicorrelated distribution $N(\mathbf{0}, \Sigma_E)$ are either 0 or .30.

All design factors were fully crossed; this yields 3 (*Data Shape*) $\times$ 3 (*Number of clusters*) $\times$ 3 (*Amount of Cluster Overlap*) $\times$ 3 (*Subspace Dimension*) $\times$ 4 (*Noise level*) $\times$ 2 (*Noise correlation*) $= 648$ combinations for each of which 10 replicates were generated, resulting in $10 \times 648 = 6,480$ simulated data sets.

All data sets will be subjected to the ALS algorithm making use of the two starting strategies as proposed in Section 3.2; in particular, we will conduct 500 runs starting from a random membership matrix (i.e., random start) and 500 runs starting from K randomly selected data points (i.e., semi-random start). Fully iterating these $1,000$ starts took on average 191 seconds.

For comparative purposes, we also included in our simulation study two sequential strategies to fit the model. The first one consists of first applying an overlapping clustering of the objects and subsequently conducting a dimensional reduction of the profiles. The second one consists of first applying a dimensional reduction of the variables of the data and subsequently fitting an additive overlapping clustering model to the scores on the reduced variables. In both approaches the clustering is performed by means of the algorithm of Section 3.2 with the profiles now being estimated using ordinary least squares regression instead of reduced rank regression (and with 500 runs starting from a random membership matrix and 500 runs starting from K randomly selected data points). In both cases, the dimensional reduction of the variables is further obtained by means of a singular value decomposition. The right singular vectors corresponding to the S largest singular values are retained as the loadings and the multiplication of the corresponding left singular vectors with the diagonal matrix containing the S largest singular values are taken as the scores. The sequential algorithms will be denoted with $CLUS$-SVD and SVD-$CLUS$, respectively; it took them on average 91 and 6 seconds to calculate their result, respectively.

As a result, an algorithmic factor of three levels is obtained. For each algorithm, the best result of its $1,000$ runs will be reported as its final result.

4.3 Evaluation Criteria

4.3.1 Minimization Measures

4.3.1.1 Finding the Optimum. For each data set we want to determine for each algorithm whether it reached the global optimum of the loss function. However, when adding noise $\mathbf{E}$ to a true matrix $\mathbf{M}$, we do not know this global optimum. Because of this we introduce the concept of *proxy* or *pseudo-optimum*. This proxy is determined for each data set separately and acts as an approximation of the global optimum. For each data set we will determine whether the best run of each algorithm reached the proxy.

In particular, the proxy for each data set is determined as follows:

1. First of all, an *upper bound* (UB) on the loss value is determined. An obvious candidate upper bound is the loss of the true underlying lowdimensional additive overlapping clustering model $\mathbf{M} = \mathbf{AP}$. However, we will use a better upper bound by running the ALS algorithm seeded with both the true memberships $\mathbf{A}$ and the conditionally optimal memberships upon the true profiles $\mathbf{P}$; the best of the two resulting loss values will be taken as the upper bound UB.
2. All three algorithms are run on the data set which yields three loss values L_1^2, L_2^2, L_3^2.
3. The value of the proxy is then given by $\min(\text{UB}, L_1^2, L_2^2, L_3^2)$.

Note that either no, one, or several algorithms can reach the proxy for a given data set at hand.

4.3.1.2 Sensitivity to Local Optima. An additional question with regard to the optimization performance of the algorithms pertains to the individual runs within the multistart procedure. We will examine for this purpose for each data set and for each algorithm the percentage of runs within the multistart procedure that reached the proxy.

4.3.2 Recovery Measures

The ALS algorithm will yield estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{P}}$. Recovery now can be measured on the level of $\mathbf{A}$ (cluster recovery), $\mathbf{P}$ (profile recovery), and on the level of the subspace itself.

4.3.2.1 Recovery of the Clustering. To evaluate the quality of the clustering $\hat{\mathbf{A}}$ found by an algorithm, we will calculate the expression

$$\left(1 - \frac{\sum_{i,k=1}^{I,K} |a_{ik} - \hat{a}_{ik}|}{IK}\right) 100. \tag{9}$$

To take the permutational freedom of the order of clusters into account, we will define the *goodness-of-clustering* (GOC) as the minimum value of (9) over all column permutations of $\hat{\mathbf{A}}$. GOC takes values in the interval $[0, 100]$, with a value of 100 meaning perfect recovery.

4.3.2.2 Profile Recovery. To evaluate the quality of the resulting profiles $\hat{\mathbf{P}}$ found by an algorithm, we will calculate the expression

$$\left(1 - \frac{\sum_{k,j=1}^{K,J} (p_{kj} - \hat{p}_{kj})^2}{\sum_{k,j=1}^{K,J} (p_{jk} - \bar{p})^2}\right) 100, \tag{10}$$

which is a rescaling of the sum of the Euclidean distances between corresponding profiles, with $\bar{p}$ denoting the mean value of all elements in $\mathbf{P}$. To take the permutational freedom of the order of clusters into account, we will define the *goodness-of-profiles* (GOP) as the minimum value of (10) over all row permutations of $\hat{\mathbf{P}}$. The GOP takes values in the interval $(-\infty, 100]$, with a value of 100 meaning perfect recovery.

4.3.2.3 Subspace Recovery. To evaluate the quality of the resulting subspace, we will calculate the angle α between the row spaces of the resulting profiles $\hat{\mathbf{P}}$ and the true profiles $\mathbf{P}$, which lies within the interval $(0, \pi/2)$; for details about the calculation, see Krzanowski (1979). We then define a *goodness-of-subspace* (GOS) measure as a rescaled version of the angle α:

$$\text{GOS} = (1 - \alpha/(\pi/2))100. \tag{11}$$

GOS takes values in the interval $[0, 100]$, with a value of 100 meaning perfect recovery and a value of 0 meaning that the true and estimated subspaces are perpendicular to each other.

4.4 Results

The average performance of the algorithms is given in Table 1. As can be seen in this table, the ALS algorithm performs satisfactory. Also, it clearly outperforms the two sequential strategies as the best run from a multistart ALS procedure more often equals the proxy, as it is less sensitive to local optima and as it shows a better recovery of the profiles.

Furthermore, considerable differences between the different cells of the simulation design were observed. To capture these more in detail, factorial analyses of variance were conducted for each recovery criterion sepa-

Table 1. Average performance of the ALS algorithm and two sequential procedures for different evaluation criteria.

	Minimization		Recovery		
	% data sets proxy reached	% multistarts reaching proxy	Clusters (GOC)	Profiles (GOP)	Subspace (GOS)
ALS	74.52	23.3	86.82	92.47	87.78
$CLUS\text{-}SVD$	20.42	7.7	84.06	83.56	86.56
$SVD\text{-}CLUS$	32.75	7.7	85.24	88.21	85.84

Table 2. Most important effects and interactions for each evaluation criterion.

Evaluation criterion	Effect / Interaction	η^2
% data sets proxy is reached on	Algorithm	.32
	Noise Level	.30
	Algorithm * Noise Level	.11
Multistarts reaching the proxy	Number of Clusters K	.09
	Noise Level	.07
Cluster Recovery (GOC)	Subspace Dimension S	.27
	Noise Level	.26
	Number of Clusters K	.08
Profile Recovery (GOP)	Noise Level	.11
Subspace Recovery (GOS)	Noise Level	.33
	Subspace Dimension S	.19

rately, with the data factors of the simulation design acting as the independent variables and the algorithmic factor being treated as repeated measures. For the criterion of whether or not the best run equals the proxy, we first calculated for each cell of the design the percentage of data sets on which the proxy had been reached by the best run and then analyzed these percentages by means of a factorial analysis of variance in which the highest order interaction term has been omitted. Below, we will only discuss main and interaction effects with an effect size η^2 (Cohen and Cohen 1983) of about .08 or higher; these are listed in Table 2.

With regard to algorithmic differences, a main effect of the algorithmic factor was found on the percentage of data sets (i.e., best run) that reached the proxy. Furthermore, for the same criterion, an interaction between the algorithmic factor and the amount of noise was observed. This interaction stemmed from the fact that the sequential algorithms show a much steeper decrease in performance for increasing amounts of noise than the simultaneous ALS algorithm.

Regarding data characteristics, first, the amount of noise influences all evaluation measures relatively strongly, with worse values showing up

for higher amounts of noise. Second, the number of clusters K has a sizeable impact on the recovery of the clustering and on the number of multistarts that find the proxy, with a larger number of clusters yielding a worse performance. Third, the dimensionality S of the subspace influences both cluster and subspace recovery. This influence, however, appears to be more complex, with a higher number of dimensions implying, on the one hand, a better cluster recovery, and, on the other hand, a worse subspace recovery.

4.5 Discussion of the Results

A main effect for the algorithmic factor was found for the criterion of the percentage of data sets that hit the proxy. This result implies that the best run of the simultaneous approach finds the proxy much more often than the best run of the sequential approaches. Moreover, as can be seen in Table 1, the simultaneous approach is less sensitive to local optima than the sequential approaches in that more multistarts reach the proxy. Furthermore, higher amounts of noise were found to have a negative influence on algorithmic performance. Also, a larger number of clusters was found to have a negative influence on the sensitivity to local optima and on the recovery of the clustering and the subspace. This finding can be attributed to the fact that the solution space for the minimization of loss function (6) grows exponentially with an increasing number of clusters K. Finally, a larger subspace dimension S appeared to have a negative influence on subspace recovery, whereas a smaller dimension appeared to have a negative influence on cluster recovery. The former effect could be due to the fact that, in case of subspace dimensionalities $S < J/2$, increasing the dimensionality S implies a more complex solution set of possible subspaces. The latter effect may be related to the observation in the simulation study of Depril et al. (2008) for the additive overlapping clustering model (1) that, if the rows of the profile matrix $\mathbf{P}$ are correlated, algorithmic performance drops; a smaller subspace dimension, indeed, can be expected to imply a larger dependence among the profiles (i.e., larger correlations among rows of $\mathbf{P}$).

To evaluate whether an algorithm found the global optimum of the loss function, we had to introduce an estimate of the global optimum as this is in general not known. For the data sets with no noise, however, the global optimum of the loss function is known (viz. the data themselves). ANOVA analyses on these noise-free data yielded the same results as the analyses of the whole of all data sets, yet, with smaller algorithmic differences and better recoveries.

The best run of the ALS algorithm was found to reach the proxy on 75% of the data sets, whereas the best run of the sequential approaches found the proxy on 20% (resp. 30%) of the data sets only (as can be seen

in Table 1). However, one could wonder whether, in addition to the ALS approach, applying one or two of the sequential approaches would yield an incremental gain in reaching the proxy. This indeed appeared to be the case for the application of the sequential SVD-$CLUS$ approach which, in combination with ALS, yielded the proxy on 88% of the data sets; no such incremental gain in reaching the proxy was obtained when combining ALS with the sequential $CLUS$-SVD approach.

To understand the latter results, one may note that the optimization problem at hand is a hard nut to crack because the large number of parameters of different types (i.e., mixed binary-continuous) that need to be estimated and the existence of many local optimal solutions; the problem of local optima is an ubiquitous challenge for almost all combinatorial optimization problems, see, for instance, Hubert, Arabie, and Hesson-McInnes (1992) and Ceulemans and Van Mechelen (2004). When looking at Table 3, in which the average values for the three algorithms for the different evaluation criteria are displayed for noise-free data (left) and data with a large amount of noise (i.e., $\varepsilon = .30$), it appears that even for noise-free data all three algorithms, although performing equally well, have problems in finding the global optimal solution (which is known in this case and equals the proxy because the data contain no noise). When more noise is added (see right part of Table 3) and thus the optimization problem becomes harder, the optimization performance decreases a bit for ALS and dramatically for the other two algorithms. However, this does not exclude the possibility that for some data sets ALS may break down and may -by accident- be outperformed by one of the other algorithms (i.e., in our case only by SVD-$CLUS$). It may be conjectured that ALS may especially break down (and possibly may be outperformed by SVD-$CLUS$) when the optimization problem becomes harder (i.e., large amounts of noise and/or a large number of clusters). When checking our results this conjecture appears to be supported.

There exist different strategies that may be applied separately or combined to alleviate the local minima problem of ALS in certain situations. First, one may increase the number of random or semi-random starts that is used. For instance, for the related *K*-means clustering problem, Steinley (2003) demonstrates that minimal $1,000$ starts (even many more in difficult situations) need to be used to arrive at a good solution. Second, one may try to improve the quality of the starting solution by determining other (better) rational starting solutions or by generating pseudo-random initial solutions (Ceulemans, Van Mechelen, and Leenen 2007). As an example of the first, the best run of the SVD-$CLUS$ algorithm may be used as a rational start for ALS. A slight perturbation of this solution would be an example of a pseudo-random start.

Table 3. Average performance of the ALS algorithm and two sequential procedures for different evaluation criteria with the Noise level ϵ being equal to 0 (left) or .30 (right).

	$\varepsilon = 0$					$\varepsilon = .30$				
	Minimization		Recovery			Minimization		Recovery		
	% data sets proxy hit	% multi-starts hit proxy	Clus-ters GOC	Pro-files GOP	Sub-space GOS	% data sets proxy hit	% multi-starts hit proxy	Clus-ters GOC	Pro-files GOP	Sub-space GOS
ALS	81.73	30.23	96.61	97.86	100.00	69.57	10.73	79.05	85.21	76.60
$CLUS\text{-}SVD$	81.73	30.61	96.25	96.31	100.00	0.00	0.00	75.94	71.48	74.47
$SVD\text{-}CLUS$	81.42	30.17	96.20	96.39	100.00	20.74	0.24	76.69	78.65	74.09

5. Application

In this section we will present an application of the lowdimensional additive overlapping clustering model to illustrate its potential usefulness. This application will be taken from the domain of the psychology of emotions. In particular, we will present an analysis of data on anger intensities as experienced by a number of persons in a number of different situations. These data can be arranged in a two-way two-mode person by situation matrix. The interest of the researcher then is to understand the variability in anger intensities across both persons and situations.

A straightforward theoretical mechanism one could appeal to explain this variability, may be based on the concept of frustration (Berkowitz 1989; Kuppens, Van Mechelen, and Smits 2003). In particular, first, each situation can be assumed to induce a certain level of frustration within each person. This frustration level may stem from one or more sources of frustration as present in the situation. Each source may induce a particular amount of frustration; the total level of frustration as induced by a situation then may be assumed to equal the sum of frustration amounts as associated with the sources it includes. Second, people can be assumed to differ in their level of sensitivity to frustration (or lack of frustration tolerance). The anger intensity of a particular person in a particular situation then finally can be assumed to equal the frustration level of the situation multiplied by the frustration sensitivity of the person.

The theory as outlined above can be formalized by means of a lowdimensional additive overlapping clustering model $\mathbf{M} = \mathbf{AP}$, with $\mathbf{P} = \mathbf{CB}'$ and with, in this particular case, dimensionality $S = 1$ (the latter implying that there is no rotational freedom for the decomposition). In particular, the grouping of the situations into overlapping clusters as represented by the situation cluster matrix $\mathbf{A}$ can be interpreted in terms of sources of frustration as present in the situations, with each cluster corresponding to a particular source of frustration. The amount of frustration as induced by

each of the frustration sources is further to be found in the column vector $\mathbf{C}$ (with the product $\mathbf{AC}$ then denoting the frustration levels as induced by the different situations). Finally, the frustration sensitivities of the persons are represented by the columns in $\mathbf{B}$. Multiplying each person's frustration sensitivity with each situation's frustration level as given in $\mathbf{AC}$ then yields the resulting anger intensities, $\mathbf{M} = \mathbf{ACB}'$.

To illustrate this in practice, we will make use of data collected by Kuppens and Van Mechelen (2007) on anger ratings on a scale from 0 (no anger) to 6 (very angry) from 357 students on 24 situations. (A description of the situations in question can be found in Appendix A). Additionally, Kuppens and Van Mechelen (2007) also administered to the same students a questionnaire measuring dispositional trait anger (Spielberger, Johnson, Russell, Crane, Jacobs, and Worden 1985). Lowdimensional additive overlapping clustering models with one dimension (i.e., $S = 1$) and the number of clusters K ranging from 1 to 6 were fitted to these data[2]. A scree plot of the proportion of explained variance for the different resulting models is presented in Figure 1. On the basis of this plot, we retained a model with $K = 3$ clusters; this model explains 55.5% of the variance in the data. When decomposing the profile matrix $\mathbf{P}$ into $\mathbf{P} = \mathbf{CB}'$, all entries of $\mathbf{B}$ and $\mathbf{C}$ appeared to be positive. We further chose the length of $\mathbf{B}$ to be such that the entries of $\mathbf{C}$ lie in the interval (0,2), implying that the frustration levels of the situations as given in $\mathbf{AC}$ lie in the interval (0,6), analogous to the persons' anger intensities.

The situation clusters are depicted as intersecting sets in Figure 2. The first cluster consists of situations in which an agreement has been broken (e.g., you arrange with a good friend to go out together, and he/she will contact you to meet each other, but you don't hear from him/her). The second cluster consists of situations in which a person suffers physical or psychological damage (e.g., a friend returns your CD player, claiming that everything is OK, but it turns out to be broken afterwards). The third cluster consists of situations with a consequential loss (e.g., on holiday with friends, you arrange that each, in turn, has to carry the heavy tent gear, but one day, the tent gear is missing).

The interpretations of the three situation clusters in terms of sources of frustration could be validated through available expert ratings of the situations. In particular, the situations belonging to the first cluster scored higher than situations not belonging to this cluster on a rating of non-compliance with an agreement (an average of 4.1 vs 2.7 on a scale from 1 to 6, $t(22) =$

2. All models were fitted with both the ALS algorithm, as proposed in Section 3.2, and the sequential approach $SVD\text{-}CLUS$, with for each algorithm 500 random memberships starts (i.e., random starts) and 500 data points starts (i.e., semi-random starts) being taken and retaining the best of these $1,000$ starts as the final estimate.

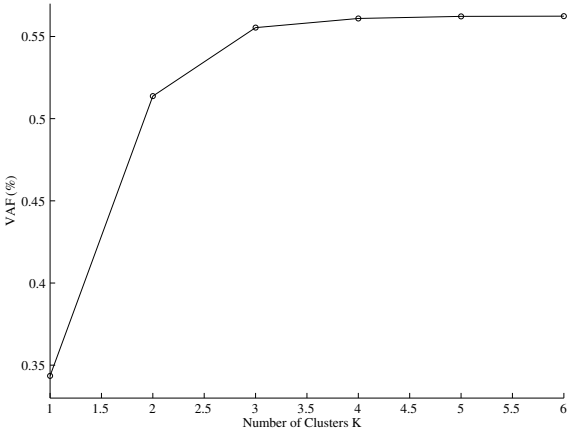


Figure 1. Scree plot for selecting a lowdimensional additive overlapping clustering model of dimension $S = 1$ for the situation by persons' anger intensities.

1.83, $p < .05$, one-tailed); situations belonging to the second cluster scored higher on induction of feelings of hurt (an average of 4.8 vs 4.3 on a scale from 1 to 6, $t(22) = 1.83$, $p < .05$, one-tailed); finally, situations of the third cluster scored higher on negative implied consequences (an average of 5.0 vs 3.9 on a scale from 1 to 6, $t(22) = 2.77$, $p < .01$, one-tailed).

The frustration levels of the three situation clusters as included in the matrix $\mathbf{C}$ take values of .5, .9, and 1.7, respectively. The order of these frustration levels somewhat reflects the importance of the obstacle as included in the situations: One may experience an unreturned love as much more awful than a damaged bike, which in turn, can be conceived as more regrettable than the cancellation of a swimming date.

The persons' frustration sensitivities are given in the matrix $\mathbf{B}$. A corresponding histogram is given in Figure 3. The average frustration sensitivity is 1.43 with a standard deviation of $.45$. The interpretation of $\mathbf{B}$ as frustration sensitivity could be externally validated in terms of a correlation of .37 ($p < .001$) with trait anger.

A graphical representation of the full lowdimensional additive overlapping clustering model with one dimension (i.e., $S = 1$) is given in Figure 4. This figure displays for each pattern of situation cluster memberships the boxplot of the individual anger intensities as predicted by the model.

6. Discussion

In this paper we proposed a novel model for two-way two-mode object by variable data to simultaneously perform an overlapping clustering

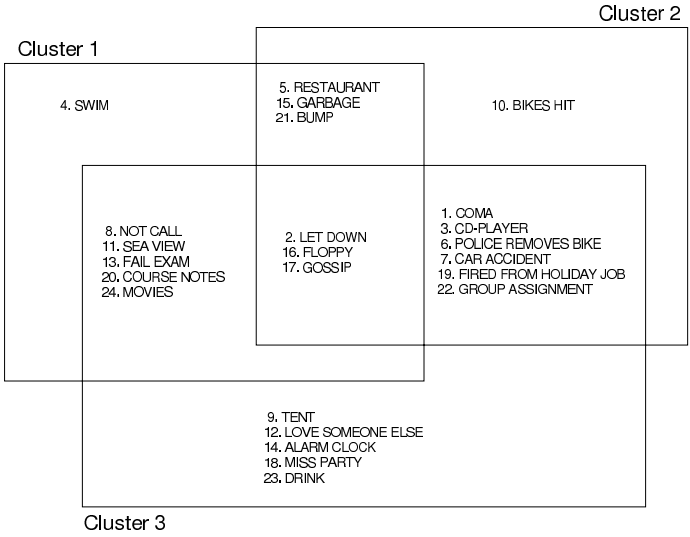


Figure 2. Graphical representation of the three situation clusters found.

of one of the modes (i.e., objects) and a dimensional reduction of the other mode (i.e., variables). This model is a constrained version of the additive overlapping clustering model proposed by Mirkin (1987), by restricting the profiles and, hence, the reconstructed data to lie in a lowdimensional subspace of the data space. Fitting this model to an empirical data set is done by minimizing a least squares loss function. To minimize the loss function we proposed an ALS algorithm. In a simulation study, it was found that the ALS algorithm works properly and that it is superior to approaches that sequentially apply a dimensional reduction of the variables and a clustering of the objects. Based on these simulation results, it may be recommended for statistical practice to combine the new ALS algorithm with the sequential $SVD\text{-}CLUS$ approach, and this especially for those optimization situations that may be expected to be very hard (i.e., a large amount of noise and a large number of underlying clusters). Finally, we illustrated the potential usefulness of the model by applying it to a data set from the domain of the psychology of emotions.

The lowdimensional additive overlapping clustering model subsumes a lowdimensional partitioning model as a special case. In the literature, two such lowdimensional partitioning models, which can be shown to be equivalent, have been proposed before: the projection pursuit partitioning model of Bock (1987) and the REDKM model of De Soete and Carroll (1994). Both models constrain the centroids of the partitioning to lie in a lowdimensional subspace of the data space. When the overlapping clustering as included in

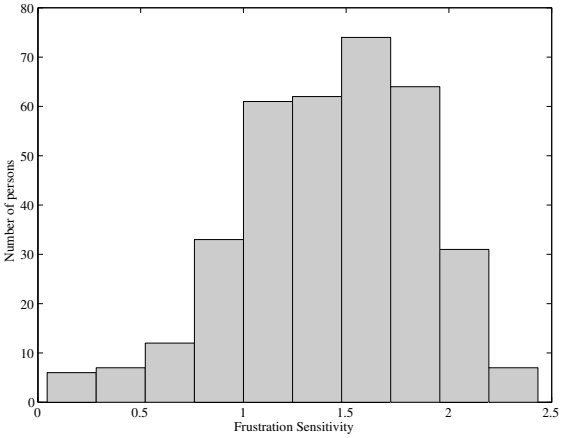


Figure 3. Histogram of the person's frustration sensitivity as given in $\mathbf{B}$.

the model that was introduced in the present paper is constrained to be a partitioning, the projection pursuit partitioning (or REDKM) model is obtained. Note, though, that these lowdimensional partitioning models have been proposed for centered data.

The S-dimensional subspace as included in the lowdimensional additive overlapping clustering model is assumed to contain the origin, which implies that the data are assumed to lie approximately in a subspace containing the origin. The question may rise whether this assumption is appropriate. If this would not be the case, one could consider either to preprocess the data prior to the actual analysis or to extend the lowdimensional additive overlapping clustering model with an offset term.

With regard to a possible preprocessing, an obvious choice would be to translate the data in such a way that the translated data subspace does contain the origin. One possibility for this could be to center the data as was proposed in the case of the lowdimensional partitioning models. One should note, however, that when considering translations, a distinction needs to be drawn between fitting a partitioning model and fitting an overlapping clustering model. We will clarify this with the following reasoning. Assume that a raw data set lies (approximately) in an S-dimensional subspace containing the origin. A researcher who would be unaware of this could nevertheless decide to center the raw data set. For a partitioning, such a centering would have no consequences for the structure of the final model; the structure of an additive overlapping clustering model, however, will generally change as a translation will affect the reconstructed data values in the cluster intersections.

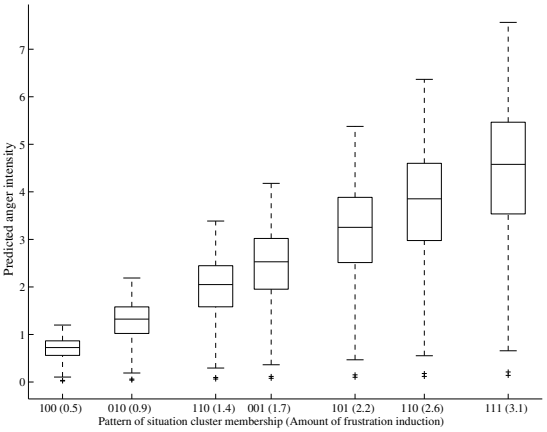


Figure 4. Boxplots of the persons' predicted anger intensities for each group of situations together with the group's level of frustration induction. The binary coding on the horizontal axis refers to the row patterns of the membership matrix. For instance "011" refers to the situations belonging to the second and third cluster.

As an alternative to preprocessing, one could incorporate an offset term in the lowdimensional additive overlapping clustering model. This, however, may imply some non-trivial identification issues that need to be taken into account.

Finally, one may remark that in the results of the analysis of the anger data in Section 5, the values of the resulting estimated matrices $\mathbf{P}$, $\mathbf{B}$, and $\mathbf{C}$ all were positive. This circumstance happened to be very useful since it allowed us to interpret the entries of these matrices as frustration intensities or sensitivities. Whereas in this example the positivity of these matrices happened 'by accident', one might wish to impose positivity constraints on some or all of the component matrices of a lowdimensional additive overlapping clustering model in an a priori way. The estimation of such a constrained model would imply an interesting challenge for further research. Otherwise, such a constrained model would line up with earlier work on non-negative matrix factorization (Lee and Seung 1999, 2001).

Appendix A. Descriptions of the Situations

Table 4 contains the descriptions and the labels of the situations of the data collected by Kuppens and Van Mechelen (2007) used in Section 5.

Table 4. Description of all twenty-four situations of the data of Kuppens and Van Mechelen (2007).

(1)	Your friend is in a coma after an accident. (COMA)
(2)	A friend lets you down on a date, and calls you the following day to let you know that he/she did not feel like meeting with you and went out with other people instead. (LET DOWN)
(3)	A friend returns your CD player, claiming that everything is OK, but it turns out to be broken afterwards. (CD-PLAYER)
(4)	A swimming appointment is canceled because one of your friends falls ill. (SWIM)
(5)	The waiter in a restaurant informs you that it may take a while before you can eat because it is a busy evening. Finally, you are served after 50 minutes of waiting. (RESTAURANT)
(6)	Upon leaving class, you notice that a police officer is removing your bike because it was illegally parked. (POLICE REMOVES BIKE)
(7)	You are hit by a car on your way to an important appointment, causing you to miss the appointment. (CAR ACCIDENT)
(8)	You arrange with a good friend to go out together, and he/she will contact you to meet each other. You don't hear from him/her. (NOT CALL)
(9)	On holiday with friends, you arrange that each, in turn, has to carry the heavy tent gear. One day, the tent gear is missing. (TENT)
(10)	You are hit on your bike by another biker. He/she apologizes, and proposes to pay back the damage to your bike. (BIKES HIT)
(11)	You have arranged for a hotel room with sea-view. Upon arrival, you are given a room without a sea-view. (SEA VIEW)
(12)	You are in love with someone but he/she is more interested in someone else. (LOVE SOMEONE ELSE)
(13)	You did not study hard enough for an exam, and you fail the exam. (FAIL EXAM)
(14)	Your clock failed to wake you up in the morning and you miss the final class of a course. (ALARM CLOCK)
(15)	You arrange with your roommates that each in turn has to put out the garbage. When it is someone else's turn you noticed that he/she did not clean up. (GARBAGE)
(16)	A floppy disk holding an important school assignment is destroyed by your computer. (FLOPPY)
(17)	You hear that a friend is spreading gossip about you. (GOSSIP)
(18)	You miss a popular party because you fall asleep at home. (MISS PARTY)
(19)	You are fired from your holiday job. (FIRED FROM HOLIDAY JOB)
(20)	A fellow student fails to return your notes when you need them for studying. (COURSE NOTES)
(21)	You bump into someone on the street. (BUMP)
(22)	You have a group assignment with some fellow students. They don't work hard, and you all get a bad grade. (GROUP ASSIGNMENT)
(23)	You're out for a drink after a hard day's work, and you have to wait 30 minutes before you are served. (DRINK)
(24)	Your roommates went to the movies without informing you. (MOVIES)

References

ANDERSON, T. (1951), "Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions," *The Annals of Mathematical Statistics 22*, 327–351.

ARABIE, P., and HUBERT, L. (1994), "Cluster Analysis in Marketing Research," in *Handbook of Marketing Research*, ed. R. Bagozzi, Oxford: Blackwell, pp. 160–189.

BERKOWITZ, L. (1989), "Frustration-aggression Hypothesis: Examination and Reformulation," *Psychological Bulletin 106*, 59–73.

BOCK, H.-H. (1987), "On the interface Between Cluster Analysis, Principal Component Analysis and Multidimensional Scaling," in *Multivariate Statistical Modeling and Data Analysis: Proceedings of the Advanced Symposium on Multivariate Modeling and Data Analysis May 15–16, 1986*, eds. H. Bozdogan and A. Gupta, Dordrecht, The Netherlands: Reidel Publishing Company, pp. 17–34.

CARROLL, J. D., and CHATURVEDI, A. (1995), "A General Approach to Clustering and Multidimensional Scaling of Two-way, Three-way or Higher-way Data," in *Geometric Representations of Perceptual Phenomena: Papers in honor of Tarow Indow on his 70th birthday*, eds. D. R. Luce, M. D'Zmura, D. Hoffman, G. J. Iverson, and K. A. Romney, Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 295–318.

CATTELL, R. B. (1966), "The Meaning and Strategic Use of Factor Analysis," in *Handbook of Multivariate Experimental Psychology*, ed. R. B. Cattell, Chicago: Rand McNally, pp. 174–243.

CEULEMANS, E., and 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.

CEULEMANS, E., TIMMERMAN, M. E., and KIERS, H. A. L. (2011), "The CHull Procedure for Selecting Among Multilevel Component Solutions," *Chemometrics and Intelligent Laboratory Systems 106*, 12–20.

CEULEMANS, E., and VAN MECHELEN, I. (2005), "Hierarchical Classes Models for Three-way Three-mode Binary Data: Interrelations and Model Selection," *Psychometrika 70*, 461–480.

CEULEMANS, E., and VAN MECHELEN, I. (2004), "Tucker2 Hierarchical Classes Analysis," *Psychometrika 69*, 375–399.

CEULEMANS, E., VAN MECHELEN, I., and LEENEN, I. (2007), "The Local Minima Problem in Hierarchical Classes Analysis: An Evaluation of a Simulated Annealing Algorithm and Various Multistart Procedures," *Psychometrika 72*, 377–391.

CEULEMANS, E., VAN MECHELEN, I., and LEENEN, I. (2003), "Tucker3 Hierarchical Classes Analysis," *Psychometrika 68*, 413–433.

CHANG, W.-C. (1983), "On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions," *Applied Statistics 32*, 267–275.

CHATURVEDI, A., and CARROLL, J. D. (1994), "An Alternating Combinatorial Optimization Approach to Fitting the INDCLUS and Generalized INDCLUS Models," *Journal of Classification 11*, 155–170.

COHEN, J., and COHEN, P. (1983), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Erlbaum.

DEPRIL, D., VAN MECHELEN, I., and MIRKIN, B. G. (2008), "Algorithms for Additive Clustering of Rectangular Data Tables," *Computational Statistics and Data Analysis 52*, 4923–4938.

DE SOETE, G., and CARROLL, J. D. (1994), "K-means Clustering an a Low-dimensional Euclidean Space," in *New Approaches in Classification and Data Analysis*, eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Berlin, Germany: Springer-Verlag, pp. 212–219.

EVERITT, B. (1977), "Cluster Analysis," in *The Analysis of Survey Data, Vol. 1: Exploring Data Structures*, eds. C. A. O'Muircheartaig and C. Payne, London: Wiley, pp. 63–88.

HUBERT, L. J., ARABIE, P., and HESSON-MCINNES, M. (1992), "Multidimensional Scaling in the City-block Metric - A Combinatorial Approach," *Journal of Classification 9*, 211–236.

KRZANOWSKI, W. (1979), "Between-groups Comparison of Principal Components," *Journal of the American Statistical Association 74*, 703–707.

KUPPENS, P., and VAN MECHELEN, I. (2007), "Determinants of the Anger Appraisals of Threatened Self-esteem, Other-blame, and Frustration," *Cognition and Emotion 21*, 56–77.

KUPPENS, P., VAN MECHELEN, I., and SMITS, D. J. M. (2003), "The Appraisal Basis of Anger: Specificity, Necessity and Sufficiency of Components," *Emotion 3*, 254–269.

LEE, D. D., and SEUNG, S. H. (2001), "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems 13*, 556–562.

LEE, D. D., and SEUNG, S. H. (1999), "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature 401*, 788–791.

MIRKIN, B. G. (1987), "Method of Principal Cluster Analysis," *Automation and Remote Control 48*, 1379–1386.

ROCCI, R., and VICHI, M. (2005), "Three-mode Component Analysis with Crisp or Fuzzy Partition of Units," *Psychometrika 70*, 715–736.

SCHEPERS, J., CEULEMANS, E., and 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.

SHEPARD, R. N., and ARABIE, P. (1979), "Additive Clustering Representations of Similarities as Combinations of Discrete Overlapping Properties," *Psychological Review 86*, 87–123.

SPIELBERGER, C. D., JOHNSON, E. H., RUSSELL, S. F., CRANE, J. C., JACOBS, G. A., and WORDEN, T. J. (1985), "The Experience and Expression of Anger: Construction and Validation of an Anger Expression Scale," in *Anger and Hostility in Cardiovascular and Behavioral Disorders*, eds. M. A. Chesney and R. H. Rosenman, New York: Hemisphere, pp. 5–30.

STEINLEY, D. (2003), "Local Optima in K-means Clustering: What You Don't Know May Hurt You," *Psychological Methods 8*, 294–304.

STEINLEY, D., and BRUSCO, M. J. (2007), "Intializing K-means Batch Clustering: A Critical Evaluation of Several Techniques," *Journal of Classification 24*, 99–121.

STOICA, P., and VIBERG, M. (1996), "Maximum Likelihood Parameter and Rank Estimation in Reduced-Rank Multivariate Linear Regressions," *IEEE Transactions on Signal Processing 44*, 3096–3078.

TRYON, R. C., and BAILY, D. E. (1970), *Cluster Analysis*, New York: McGraw-Hill.

VICHI, M., and KIERS, H. A. L. (2001), "Factorial K-means Analysis for Two-Way Data," *Computational Statistics and Data Analysis 37*, 49–64.

VICHI, M., ROCCI, R., and KIERS, H. A. L. (2007), "Simultaneous Component and Clustering Models for Three-Way Data: Within and Between Approaches," *Journal of Classification 24*, 71–98.

WILDERJANS, T. F., CEULEMANS, E., and KUPPENS, P. (2012), "Clusterwise HICLAS: A Generic Modeling Strategy to Trace Similarities and Differences in Multi-Block Binary Data," *Behavior Research Methods, 44*, 532–545.

WILDERJANS, T. F., CEULEMANS, E., and MEERS, K. (in press), "CHull: A Generic Convex Hull Based Model Selection Method," *Behavior Research Methods*.

WILDERJANS, T. F., CEULEMANS, E., and VAN MECHELEN, I. (in press), "The SIMCLAS Model: Simultaneous Analysis of Coupled Binary Data Matrices with Noise Heterogeneity Between and Within Data Blocks," *Psychometrika*.

WILDERJANS, T. F., CEULEMANS, E., VAN MECHELEN, I., and DEPRIL, D. (2011), "ADPROCLUS: A Graphical User Interface for Fitting Additive Profile Clustering Models to Object by Variable Data Matrices," *Behavior Research Methods 43*, 56–65.