Chemometrics and Intelligent Laboratory Systems 154 (2016) 194–206

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Two-mode K-spectral centroid analysis for studying multivariate longitudinal profiles☆

Joke Heylen, Iven Van Mechelen, Eiko I. Fried, Eva Ceulemans *

KU Leuven, Research Group of Quantitative Psychology and Individual Differences, Tiensestraat 102-bus 3713, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form 3 March 2016
Accepted 8 March 2016
Available online 17 March 2016

Keywords:
K-spectral centroid analysis
Biclustering
Time profiles
Multiway data

ABSTRACT

In many scientific areas, researchers collect multivariate time profile data on the evolution of a set of variables across time for multiple persons. For instance, clinical studies often focus on the effects of an intervention on different symptoms for multiple persons, by repeatedly measuring symptom severity for each symptom and each person. To pursue an insightful overview on how these time profiles vary as a function of both symptoms and persons, we propose two-mode K-Spectral Centroid (2M-KSC) analysis, which is a multivariate extension of K-Spectral Centroid analysis. Specifically, 2M-KSC assigns the persons to a few person clusters and the symptoms to a few symptom clusters and imposes that the time profiles that correspond to a specific combination of a person cluster and a symptom cluster have the same shape, but may vary in amplitude scaling. An algorithm for fitting 2M-KSC is proposed and evaluated in a simulation study. Finally, the new method is applied to time profiles regarding the severity of depression symptoms during a citalopram treatment.

1. Introduction

In many research areas interest in understanding how multiple variables change over time increases. Good examples, on which we focus in this paper, are intervention studies, targeting specific medical or psychological problems (e.g., [1,2]). In such studies, one often measures the evolution of multiple symptoms for multiple persons at consecutive time points, where a score of zero reflects the absence of a symptom. For instance, the NIH-supported STAR*D study (data version 3.0) [3,4], which we will revisit in this paper, mapped the severity of fourteen depressive symptoms for clients with major depressive disorder (MDD), and receiving a citalopram treatment, across several weeks.

The evaluation of such time profiles allows the researcher to address several important questions, including: How fast does the effect of the intervention kick in for different symptoms? Do relapses occur for some symptoms? When is full effect of the treatment reached? Are these effects across time the same for the different symptoms or can a few symptom groups be discerned each reacting differently? Similarly, what about individual differences: Is there evidence that the shape of the time profiles depends on the persons involved, and, if so, which types of persons react similarly? Obviously, these questions are important, since they allow predicting how a specific individual with a particular symptom profile would react to the intervention under study.

To further clarify these research questions and the associated modeling challenges, it is instructive to briefly review the different modeling approaches for analyzing time profiles. To this end, it is useful to distinguish between three modeling levels (for similar distinctions, see [5,6]): the *phenotype level*, the *constituent level*, and the *generating level.* Approaches at the *phenotype level* model examine which time profiles have the same manifest appearance, for instance, by clustering them into a few types. Approaches at this level differ in which profile characteristics are taken into account or sidelined when deciding whether profiles have the same shape or not. Specifically, one may take timing differences (i.e., phase variability, [7,8]) between the time profiles into account, by deciding that profiles that are time shifted (complete profile is shifted by a few time points) or warped (compressing some parts of the profile while stretching out others) versions of another differ in shape. If such differences are sidelined, however, they are removed before conducting the shape comparison. This implies that within each type, room is left for heterogeneity with respect to differences in these profile characteristics. The same holds for intensity differences (i.e., amplitude variability) between the time profiles, such as intensity shifting (complete profile is shifted in intensity by adding a scalar) or amplitude scaling (complete profile is deflated or inflated by multiplying it with a scalar).

At the *constituent level* approaches focus on the underlying constituents or components of the time profiles. For example, growth curve and trajectory models (e.g., [9,10,11]) can be situated at this level, as they model time profiles as a weighted sum of linear, quadratic, etc. basis functions and thus summarize the profiles in terms of intercepts and

☆ Selected papers from ThRee-way methods In Chemistry And Psychology conference, 7th edition, 2015.

* Corresponding author. Tel.: +32 16 325881.
E-mail addresses: joke.heylen@kuleuven.be (J. Heylen), iven.vanmechelen@kuleuven.be (I. Van Mechelen), eiko.fried@kuleuven.be (E.I. Fried), eva.ceulemans@kuleuven.be (E. Ceulemans).

http://dx.doi.org/10.1016/j.chemolab.2016.03.009
0169-7439/© 2016 Elsevier B.V. All rights reserved.

slopes. Another example of a *constituent level* method is the method of Heard, Holmes, Stephens, Hand and Dimopoulos [12], as it models time profiles as weighted combinations of prespecified nonlinear basis functions and clusters profiles based on these weights.

At the *generating level*, one is interested in the mechanism that generates the time profiles and aims to discover the underlying laws. For instance, approaches using differential equations, that relate observed scores on a variable (e.g., symptom severity) to its rate of change (e.g. [13]), or Markov approaches, where the present state of a variable (e.g., symptom severity) is dependent on the immediately preceding state only (e.g. [14]), fall within this level.

If we return to our research questions – does the shape of the time profiles vary as a function of the persons and symptoms under study, and can person types and symptom groups be induced that have similar time profiles –, it is clear that the resulting modeling challenges pertain to the manifest appearance of the profiles and thus are located at the *phenotype level.* Moreover, in the case of symptom profiles, timing differences should be taken into account when categorizing the profiles as similar or not, since they can be meaningfully interpreted as delayed or accelerated reactivity to the intervention. In addition, vertical profile or severity shifts should be taken into account as well, not in the least because zero severity values have a clear meaning (viz., symptom absence), which disappears after an upward profile shift. Differences in amplitude scaling can be sidelined, however, because they might be due to differences in the overall severity of the symptoms or their wording (e.g., suicidal thoughts vs. waking up too early) or to interindividual differences in response style; note that such scaling differences do not affect zero scores.

Among the existing approaches at the phenotype level, the method that most closely meets our modeling needs is K-Spectral Centroid (KSC) analysis ([15]; for an application in emotion psychology, see [16]). KSC clusters time profiles based on their shape, while allowing for amplitude scaling differences among the profiles that belong to the same cluster. However, KSC is a univariate method, in that it models differences in the time profiles of one symptom, or one variable in general. Therefore, the aim of this paper is to develop a multivariate extension of KSC, called two-mode KSC (2M-KSC), that allows modeling how time profiles vary as a function of both the persons and the symptoms under study. Specifically, 2M-KSC assigns the persons to a few persons clusters and the symptoms to a few symptom clusters and imposes that the time profiles that correspond to a specific combination of a person cluster and a symptom cluster have the same shape, but may vary in amplitude scaling.

The remainder of this paper is organized as follows: In the next section, the new 2M-KSC model is introduced. In Section 3, we discuss the 2M-KSC loss function and an algorithm for estimating the model parameters. Next, we elaborate on model selection. Section 4 reports a simulation study to evaluate the performance of this algorithm. In Section 5 we apply 2M-KSC to dataset version 3.0 of the STAR*D study. Finally, in Section 6, we demonstrate the usefulness of 2M-KSC in other domains of application and compare our method with existing, related phenotype methods.

2. Model

As stated above, 2M-KSC is a model for multivariate time profiles. More specifically, 2M-KSC assumes that J symptoms are measured at T time points for I persons. The T time points are comparable across the persons and the symptoms, implying that the data can be meaningfully arranged in a three-way three-mode data array $\underline{\mathbf{X}}$. Throughout this subsection we will make use of the hypothetical data set in Fig. 1, which consists of time profiles of the day-to-day severity of 4 depression symptoms collected for five MDD persons, across 10 treatment days. This data set can be perfectly reconstructed by a 2M-KSC model.

2M-KSC simultaneously clusters the I persons into K person clusters and the J symptoms into C symptom clusters. This clustering is exclusively based on the shape of the time profiles under study, discarding any amplitude scaling differences (while taking into account time shifts, warps, and severity shifts, see Introduction). All the time profiles that correspond to a specific combination of a person cluster and a symptom cluster are modeled with one particular reference profile, which reflects their typical evolution over time. Furthermore, each observed time profile receives an amplitude score, indicating its overall intensity relative to its corresponding reference profile. Specifically, this amplitude score indicates how much the reference profile has to be inflated or deflated to obtain the observed profile.

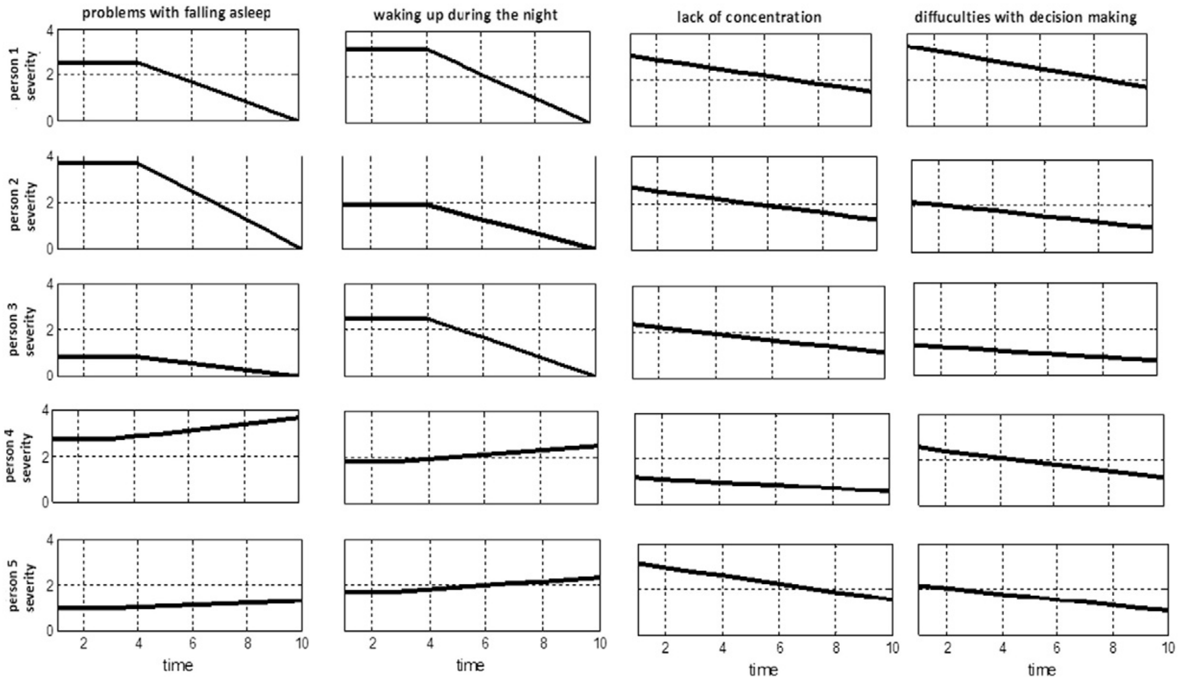

Fig. 1. Hypothetical time profiles of four depression symptoms for five MDD persons across ten treatment days.

Table 1
Partition scores p_{ik} and p_{jc}, and amplitude scores f_{ij} of the 2M-KSC model with two person and two symptom clusters, for the hypothetical dataset in Fig. 1.

Person partition scores			Symptom partition scores			Amplitude scores				
Person	Cluster 1	Cluster 2	Symptom	Cluster 1	Cluster 2		Problems falling asleep	Waking up during night	Lack of concentration	Difficulties decision making
Person 1	1	0	Problems falling asleep	1	0	Person 1	5.9	7.4	7.2	8.2
Person 2	1	0	Waking up during night	1	0	Person 2	8.5	4.4	6.4	5.2
Person 3	1	0	Lack of concentration	0	1	Person 3	1.9	5.8	5.6	3.1
Person 4	0	1	Difficulties decision making	0	1	Person 4	9.8	6.7	2.9	6.1
Person 5	0	1				Person 5	3.4	6.3	7.5	5.2

More formally, each observed time profile $\mathbf{x}_{ij}$ $(T\text{x}1)$ is modeled[1] as:

$$\mathbf{x}_{ij} = \sum_{k=1}^{K}\sum_{c=1}^{C} p_{ik}p_{jc}f_{ij}\mathbf{b}_{kc} + \mathbf{e}_{ij} \tag{1}$$

with p_{ik} a binary partition class membership score indicating to which of the K person clusters the ith person belongs (where each person belongs to a single cluster only), p_{jc} a binary partition class membership score indicating to which of the C symptom clusters the jth symptom belongs (again, each symptom belongs to a single cluster only), f_{ij} the amplitude score of the time profile of symptom j and person i, $\mathbf{b}_{kc}$ $(T\text{x}1)$ the reference profile of the bicluster resulting from the combination of person cluster k and symptom cluster c, and $\mathbf{e}_{ij}$ $(T\text{x}1)$ the residual scores. To identify the obtained solution, the reference profiles $\mathbf{b}_{kc}$ are scaled to a norm of one. Like Heylen, Van Mechelen, Verduyn, and Ceulemans [21], we discard the profile alignment feature of the KSC model of Yang and Leskovec [15] as time shifts between symptom severity profiles are meaningful shape differences.

For instance, Table 1 and Fig. 2 depict a 2M-KSC model of the hypothetical data set in Fig. 1. From Table 1, we can distinguish two person clusters, the first consisting of the first three persons and the second of persons 4 and 5. Moreover, the model contains two symptom clusters, the first containing symptoms that relate to sleep problems (problems falling asleep, waking up during the night) and the second symptoms that relate to cognitive difficulties (lack of concentration, difficulties with decision making). As can be seen from Fig. 2, which displays the time profiles that are assigned to each bicluster as well as the corresponding reference profile multiplied with the mean amplitude score for the corresponding bicluster, the two person clusters differ strongly with respect to the evolution of their sleep symptoms (symptoms improve for person cluster 1 after the fourth treatment day, but worsen for person cluster 2), whereas for cognitive symptoms differences are small (steady improvement). However, the severity of the symptoms largely differs within each bicluster, as is also revealed by the amplitude scores in Table 1.

3. Data analysis

3.1. Loss function

For a given number of person clusters K and symptom clusters C, the aim of a 2M-KSC analysis is to find the person and symptom partitions, amplitude coefficients f_{ij} and reference profiles $\mathbf{b}_{kc}$ that minimize the following least squares loss function:

$$L = \sum_{i=1}^{I}\sum_{j=1}^{J}\left\| \mathbf{x}_{ij} - \sum_{k=1}^{K}\sum_{c=1}^{C} p_{ik}p_{jc}f_{ij}\mathbf{b}_{kc} \right\|^2. \tag{2}$$

3.2. Algorithm

To minimize loss function (2), we propose to use an alternating least square (ALS) algorithm, which consists of the following steps:

1. Initialize the person and symptom partitions: Randomly assign the I persons to the K person clusters, and the J symptoms to the C symptom clusters. Each person cluster and each symptom cluster have an equal probability of being assigned to. No empty clusters are allowed.
2. For each combination of a person and symptom cluster, estimate the corresponding reference profile $\mathbf{b}_{kc}$ and amplitude coefficients f_{ij}: For each combination of a person cluster k and a symptom cluster c (i.e., bicluster$_{kc}$) collect the I_{kc} profiles assigned to this bicluster in a matrix $\mathbf{X}_{kc}$ $(T\text{x}I_{kc})$. Conduct an eigenvalue decomposition on $\mathbf{X}_{kc}\mathbf{X}'_{kc}$.[2] The eigenvector that corresponds to the largest eigenvalue is used as the estimate of the reference profile, while the amplitude coefficients f_{ij} of the I_{kc} time profiles are computed as: $f_{ij} = \mathbf{x}'_{ij}\mathbf{b}_{kc}(\mathbf{b}_{kc}'\mathbf{b}_{kc}) = \mathbf{x}'_{ij}\mathbf{b}_{kc}$. Note that this step is equivalent to extracting the first principal component from $\mathbf{X}_{kc}$. This boils down to computing the singular value decomposition of $\mathbf{X}_{kc}$ and setting $\mathbf{b}_{kc}$ to the first left singular vector and the f_{ij} scores to the corresponding entries of the first right singular vector times the first singular value.
3. For each person i, update the optimal assignment to a person cluster k, given the current estimates of p_{jc} and $\mathbf{b}_{kc}$: For this purpose, two steps are taken, conditional upon the current estimates of the symptom partition and the reference profiles. First, we compute the amplitude coefficient $f_{ij}^{(k)}$ for all J time profiles of person i if that person would be re-assigned to cluster k (for each of the K person clusters; note that k determines which reference profile is used): $f_{ij}^{(k)} = \mathbf{x}'_{ij}\mathbf{b}_{kc}$. This yields a K x J amplitude coefficients matrix for person i. Second, we calculate for each person cluster k the contribution of person i to the overall loss function (given the current parameter estimates) if that person would be re-assigned to that cluster, $\mathrm{L}_{ik} = \sum_{j=1}^{J}\left\| \mathbf{x}_{ij} - \sum_{c=1}^{C} p_{jc}f_{ij}^{(k)}\mathbf{b}_{kc} \right\|^2$, selecting the appropriate J amplitude coefficients. Person i subsequently is assigned to the person cluster k for which L_{ik} is minimal. Check whether each of the person clusters contains at least one person. If this is not the case, move the person that fits its current cluster the least to the empty cluster[3].

[1] The 2M-KSC decomposition rule bears some resemblance to that of the Tucker2 [17,18] and Tucker2-HICLAS [19,20] models. The difference with Tucker2 is that p_{ik} and p_{jc} are binary partitioning scores rather than continuous values, strictly imposing simple structure on the person and symptom modes, and that the 2M-KSC decomposition, unlike that of Tucker2, includes f_{ij} values to capture amplitude scaling differences. 2M-KSC differs from Tucker2-HICLAS in that the latter model is based on an overlapping clustering of the person and symptom modes and is intended for modeling binary data, for which amplitude scaling is no issue.

[2] Note that Yang and Leskovec [15] perform an eigenvalue decomposition on $\mathbf{X}_{\boldsymbol{k}c}\mathbf{X}_{\boldsymbol{k}c}{}' - (I_{kc}\mathbf{I}(T))$, where $\mathbf{I}$ indicates a T x T identity matrix and T reflects the number of time points. However, it can be proven making use of properties of the eigenvalue decomposition, that decomposing $\mathbf{X}_{kc}\mathbf{X}_{\boldsymbol{k}c}{}'$ leads to precisely the same eigenvectors.

[3] Before moving this subject, we check whether this subject is the only subject in its current subject cluster. If this is the case we move on to the subject that fits its current cluster the second least, and so on.

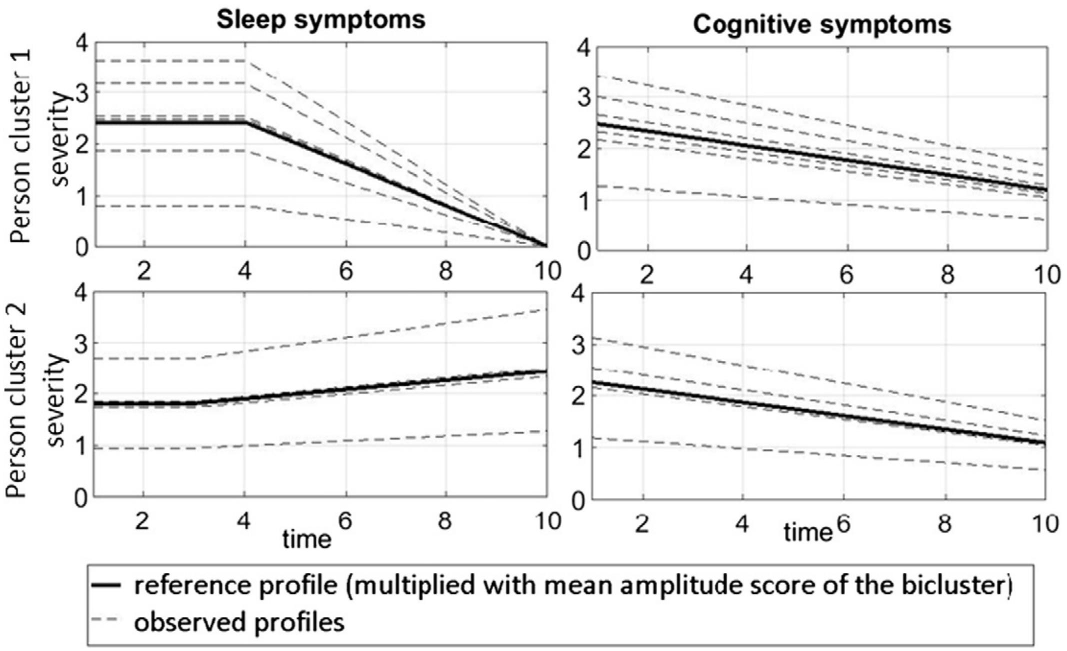

Fig. 2. Reference profiles multiplied with the mean amplitude score of the corresponding bicluster of the 2M-KSC model with two person and two symptom clusters for the hypothetical data in Fig. 1. The observed time profiles are displayed as well.

4. Update the reference profiles $\mathbf{b}_{kc}$ and amplitude coefficients by executing Step 2.
5. For each symptom j, update the optimal assignment to a symptom cluster c, given the current estimates of p_{ik} and $\mathbf{b}_{kc}$: Exchanging the roles of the persons and the symptoms, this step is completely analogous to Step 3.
6. Update the reference profiles $\mathbf{b}_{kc}$ and amplitude coefficients by executing Step 2.
7. Repeat Steps 3 to 6 until the algorithm has converged, that is, until the decrease in loss function L (2) is smaller than 10^{-6}.

Note that in the original KSC algorithm of Yang and Leskovec [15], Step 2 is preceded by scaling each time profile to a norm of one. This scaling is done to ensure that each time profile has a similar impact on the shape of the reference profile. We discarded it in our algorithm, since it also has an important drawback: It increases the weight of symptoms that hardly occur.

As generally holds for ALS algorithms (see e.g., [22,23]), the 2M-KSC algorithm may end in a local minimum. Therefore, we propose to use a multistart procedure in which the algorithm is run a large number of times (e.g., 500 and preferably more, if computation time allows) using a different random initialization of both the person and symptom partition. Also, the use of a rational start for the patient and symptom partitions can be considered. To this end, we propose to conduct a three-mode partitioning [24] with 500 random initializations of both the person and symptom partitions. Being a three-mode extension of K-means clustering, three-mode partitioning clusters the elements of each mode into mutually exclusive groups and summarizes the interrelations between these three sets of clusters in a so-called core array, of which the entries equals the means of the corresponding data scores. Here, we extract K clusters for the person mode, C clusters for the symptom mode and T clusters for the time mode; the latter actually implies that the time mode is clustered into a trivial partition of singleton classes. Thus, the core array vector that corresponds to the cth symptom cluster and kth person cluster yields an estimate of the reference profile of bicluster$_{kc}$. Note that, before conducting the three-mode clustering, we scale the time profiles to a norm of one. This scaling discards all amplitude scaling differences implying that the method will focus on shape differences. The scaling step is necessary when using the K-means extension because, unlike KSC, K-means based methods will entangle shape and amplitude when clustering time profiles, since these methods do not leave room for amplitude scaling differences within a cluster.

3.3. Model selection

For empirical data sets, the optimal number of person clusters K and/or symptom clusters C is usually unknown. The resulting model selection problem can be tackled by estimating 2M-KSC solutions with 1 up to K^{max} person clusters and 1 up to C^{max} symptom clusters, and retaining a model that has a good balance between fit and complexity, and is stable as well. To this end, we recommend to combine the CHull procedure of Ceulemans and Kiers ([25,26]; for software, see [27]) with a split-half stability analysis. Of course, substantive reasoning may further support the final choice for a specific solution.

The CHull procedure generalizes the well-known scree test of Cattell [28]. Based on a complexity (X-axis) versus fit (Y-axis) plot of the different (K, C) solutions, it looks for the solution for which holds that the gain in fit due to additional person or symptom clusters levels off. As a measure of fit we will use the percentage of the sum of the squared data entries accounted for by the model. Regarding complexity, we propose to use the sum $K + C$ of the number of person and symptom clusters (see [29]). CHull first selects the solutions with a good fit-complexity balance by discarding all solutions that are located below the higher boundary of the convex hull of the complexity versus fit plot. Next, it computes the scree test ratio's st_h of the retained (so-called) hull solutions:

$$st_h = \frac{f_h - f_{h-1}}{c_h - c_{h-1}} \Bigg/ \frac{f_{h+1} - f_h}{c_{h+1} - c_h} \qquad (3)$$

with f_h and c_h the fit and complexity values of the hth hull solution respectively, to determine how much fit increases by allowing for additional clusters. It is recommended to pick a solution with a high scree test ratio.

To evaluate the split-half stability of a specific solution, one should first decide which data mode (i.e., persons or symptoms) can be considered the sampling mode, containing a random sample of the population under study. Note that the time mode is always left intact since the shape of the time profiles refers to the intensity scores on all time points. Usually, the sampling mode will be the person mode. Provided that the sampling mode is sufficiently large, we will randomly split it into two halves (while leaving the two other data modes intact). Next, the analysis is re-run on the two halves and stability is evaluated by comparing the obtained partitions and reference profiles for the two halves to those of the original solution.

4. Simulation study

4.1. Design and procedure

The aim of this simulation study is to evaluate the performance of the proposed two-mode KSC algorithm in finding the best-fitting solution and recovering the true underlying model. In this study the number of persons I was fixed to 40 and the number of symptoms J to 16. Seven other data characteristics, that are expected to influence the performance of the algorithm [30,31,32], were manipulated in a full factorial design:

1. the number of time points T at two levels: 5 and 20;
2. the number of person clusters K at 2 levels: 2 and 4;
3. the number of symptom clusters C at 2 levels: 2 and 4;
4. the size of the person clusters at 3 levels [33]: equal, unequal with majority (60% of persons in one person cluster, the other persons equally distributed over the remaining person clusters), unequal with minority (10% of the persons in one person cluster, the other persons equally distributed over the remaining person clusters);
5. the size of the symptom clusters at 3 levels [33]: equal, unequal with majority, unequal with minority. This is done in the same way as described above;
6. the lowest congruence between the reference profiles that are associated with two person (resp. symptom) clusters, conditional upon a specific symptom (resp. person) cluster, at two levels: low (minimal ϕ between 0.00 and 0.50) and high (minimal ϕ between 0.70 and 0.90);
7. the amount of error e, which is the proportion $\frac{\|\mathbf{E}\|^2}{\|\mathbf{X}\|^2}$ of the expected sum of the squared residuals in the residual matrix $\mathbf{E}$ and the expected sum of the squared observations in the data matrix $\mathbf{X}$, at 3 levels: 0.20, 0.40, and 0.60.

We generated 20 data matrices for each cell of the design, by constructing the time profile of person i and symptom j as follows:

$$\begin{aligned}\mathbf{x}_{ij} &= \mathbf{t}_{ij} + \mathbf{e}_{\boldsymbol{ij}.}\\ &= f_{ij}^{\text{true}}\mathbf{b}_{kc}^{\text{true}} + \mathbf{e}_{ij}\end{aligned}$$

where $\mathbf{t}_{ij}$ is the true underlying time profile, resulting from the true amplitude score f_{ij}^{true} and the true reference profile $\mathbf{b}_{kc}^{\text{true}}$, which is associated with the combination of person cluster k, to which person i is allocated and symptom cluster c, to which symptom j is assigned. The true amplitude scores f_{ij}^{true} were randomly sampled from the normal distribution $N(50,10)$ (truncated at 0). The partition scores p_{ik} and p_{jc} were generated by first computing the size of the different person clusters and symptom clusters (i.e., according to the design) and then randomly assigning the correct number of persons and symptoms to the clusters. The reference profiles $\mathbf{b}_{kc}^{\text{true}}$ were generated as the weighted combination of three different probability density functions (pdf), namely beta, lognormal and normal pdfs. The weight of the first pdf (w_1) was uniformly sampled between 0 and 100, the weight of the second pdf (w_2) between 0 and 100-w_1 and the weight of the third pdf (w_3) amounted to $100 - (\mathrm{w}_1 + \mathrm{w}_2)$. The parameters of the beta pdf were uniformly sampled between 1 and 10.5, the lognormal mean e^{μ} between 0 and T and the lognormal standard deviation between 0 and $T/5$, and both the normal mean and standard deviation between 0 and T. Next, we ran through all combinations of two different person clusters and computed, for each symptom cluster, the Tucker congruence between the associated reference profiles. Furthermore, we did the same for all combinations of two different symptom clusters, where for each person cluster the congruence of the reference profiles was computed. At this point, it was checked whether the minimal congruence of all these combinations met the specified criterion, i.e., between .00 and .50 in the low congruence conditions and between .70 and .90 in the high congruence conditions; if not, we generated new reference profiles. To illustrate this further, Fig. 3 displays the true underlying reference profiles of two simulated datasets, one with low minimal congruence (Fig. 3a) and one with high minimal congruence (Fig. 3b). Both datasets have 20 time points, two person clusters, and two symptom clusters. Specifically, for the low minimal congruence data set this means that: (1) within at least one of the two person

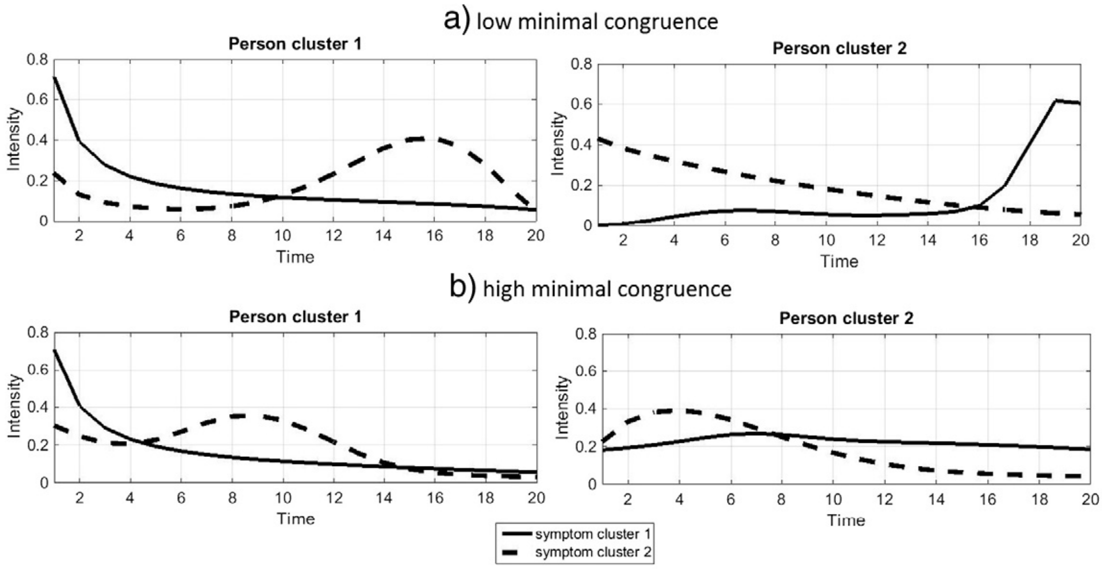

Fig. 3. True underlying reference profiles for two simulated data sets with 20 time points, two person clusters, and two symptom clusters, and with low minimal congruence (a) and high minimal congruence (b).

clusters, the congruence between the reference profiles of the two symptom clusters is low (i.e., between .00 and .50: within person cluster 1, $\phi = .57$; within person cluster 2: $\phi = .28$), and (2) within at least one of the two symptom clusters, the congruence between the reference profiles of the two person clusters is low as well (within symptom cluster 1, $\phi = .23$; within symptom cluster 2 $\phi = .60$). For the high minimal congruence data set: (1) across the person clusters, the minimal congruence between the reference profiles of the two symptom clusters is high (i.e., between .70 and .90: within person cluster 1, $\phi = .88$; within person cluster 2 $\phi = .85$), and (2) across the symptom clusters, the minimal congruence between the reference profiles of the two person clusters is high as well (within symptom cluster 1, $\phi = .71$; within symptom cluster 2 $\phi = .87$).

As we focus on symptom severity time profiles in this paper, which are always non-negative, the residuals e_{ij} were sampled from a truncated normal distribution $N(\mu,\sigma_{ij}^2)|e_{ij}\geq -t_{ij}$, with $\mu = 0$ and σ_{ij}^2 chosen in such a way that $\sum_{i=1}^{I}\sum_{j=1}^{J}\mathrm{E}(e_{ij}{}^2) = \|\mathbf{T}\|^2\frac{e}{1-e}$, with $\mathbf{T}$ the true data matrix.

In total, 2 (number of time points) × 2 (number of person clusters) × 2 (number of symptom clusters) × 3 (size of person clusters) × 3 3 (size of symptom clusters) × 2 (congruence of the reference profiles) × 3 (amount of error) × 20 (replicates) = 8640 simulated data sets were generated. Each data set was analyzed with 2M-KSC, using the correct number of person clusters K and symptom clusters C, using 500 random initializations and one rational initialization of the partition scores p_{ik} and p_{jc}, as described in the previous section.

4.2. Results

4.2.1. Sensitivity to local minima

Ideally, we want our algorithm to find the global minimum, that is, the solution associated with the lowest possible loss function value. However, due to the error perturbations in our simulated data, this global minimum is unknown (see e.g., [34]). Therefore, we obtain a surrogate global minimum by seeding the 2M-KSC algorithm with the correct partition scores p_{ik} and p_{jc}. Subsequently, we compared the loss function value of the best solution out of the 501 runs (500 random and one rational) with this surrogate global minimum value. If the difference between the loss function value of the best solution and the loss function value of the surrogate global minimum value is bigger than 10^{-6}, our algorithm ended in a local minimum for sure. This was the case for 369 out of the 8640 data sets (4.3%). The majority of the associated data sets contained four symptom clusters (72%), four subject clusters (68%), had a high degree of congruency (66%), and/or contained a lot of error (i.e., $e = .60$; 54%).

For each data set we computed the attraction rate, which equals the percentage of the 501 initializations that ends up in a loss function value that differs less than 10^{-6} from that of the finally selected solution. On average, we found an attraction rate of 34% (171 out of 501 starts). To analyze how the attraction rate differs as a function of the manipulated characteristics, we performed an analysis of variance (ANOVA). Only considering the effects for which the partial eta-squared values η_p^2 exceed .20, we found sizeable main effects of the number of person clusters ($\eta_p^2 = .45$), the number of symptom clusters ($\eta_p^2 = .31$), the size of the symptom clusters ($\eta_p^2 = .21$), the degree of congruence between the reference profiles ($\eta_p^2 = .40$), and the amount of error on the data ($\eta_p^2 = .56$). In general, less random runs end up in the retained solution when the data are more complex, the reference profiles are more congruent and the data contain more error (see Fig. 4). Since all of these data characteristics are almost always unknown beforehand when analyzing an empirical data set, and since tracking attraction rates during the analysis is cumbersome (e.g., whenever a better run is encountered, the attraction rate has to be set to zero again), we advise to consistently use a high number of random starts when running 2M-KSC analyses.

To examine if the attraction rate predicts whether a data set ends up in a local minimum, we performed a logistic regression with a binary dummy, indicating whether the analysis yielded a local minimum for sure, as dependent variable and the attraction rate as independent variable. The results revealed that a lower attraction rate (%) significantly contributes to the odds of ending up in a local minimum ($B = -0.18$, $p < .005$). Based on this result, we recommend to rerun the analysis whenever one encounters low attraction rates, to double-check the findings.

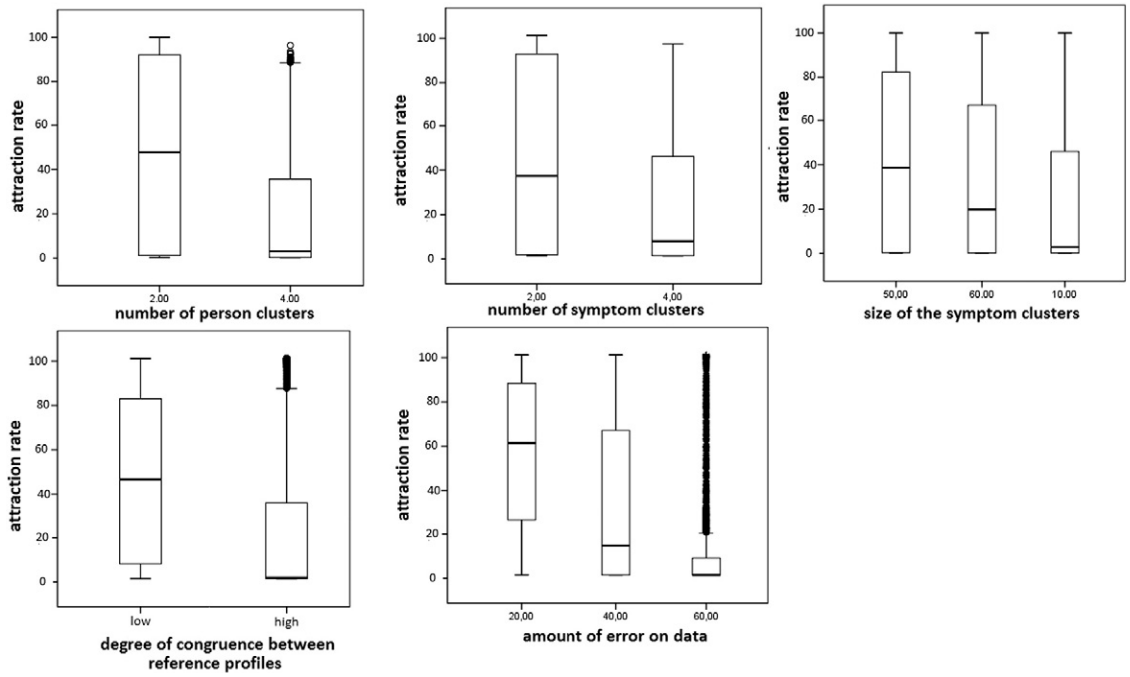

Fig. 4. Boxplots of the attraction rates (%) as a function of the number of person clusters, the number of symptom clusters, the size of the symptom clusters, the degree of congruence between reference profiles, and the amount of error on the data.

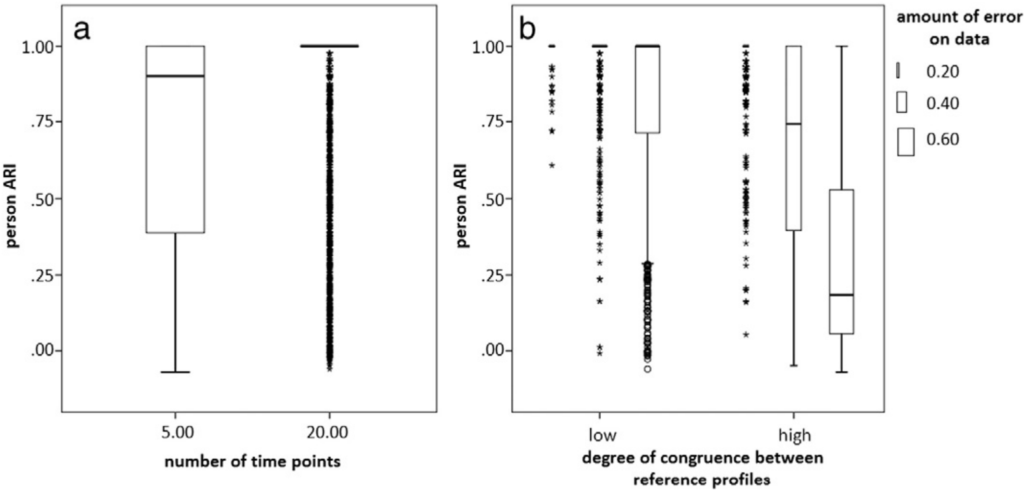

Fig. 5. Boxplots of the person ARI as a function of (a) the number of time points and (b) the congruence between reference profiles and the amount of error on the data.

4.2.2. Goodness of recovery

We evaluated the goodness of recovery of the obtained solutions with respect to (a) the person clustering and (b) the symptom clustering.

(a) Recovery of the person clustering

We used the Adjusted Rand Index (ARI; [35]), computed between the true person partition and the estimated one, to examine how well the person clustering was recovered. This ARI value is one when both partitions are equal, zero when they resemble each other as expected by chance and is negative when they resemble each other less than expected by chance.

Overall a mean person ARI of .81 was found. Moreover, for the majority of data sets (5479 out of 8640; 63%) the person ARI equals one. We performed an ANOVA, with the person ARI as dependent variable and the seven manipulated data characteristics as independent variables to investigate their influence on the person ARI. We found sizeable main effects of the number of time points ($\eta_p^2 = .35$), the degree of congruence between reference profiles ($\eta_p^2 = .41$), and the amount of error on the data ($\eta_p^2 = .53$), and an interaction effect of congruence between reference profiles and amount of error ($\eta_p^2 = .26$). These effects imply that the person clustering is recovered worse when less time points are available, when the data contain more error and when reference profiles are more congruent, and that the congruence effect is stronger in case the error level is higher (see Fig. 5).

(b) Recovery of the symptom clustering

We found an overall mean symptom ARI of .87. Furthermore, for 6821 out of the 8640 data sets (80%) a symptom ARI value of one is found. We studied the influence of the manipulated characteristics on the variable ARI by means of an ANOVA. Sizeable main effects of the number of time points ($\eta_p^2 = .23$),the congruence between reference profiles ($\eta_p^2 = .26$), and the amount of error on the data ($\eta_p^2 = .38$) were found. These effects imply that the symptom clustering is recovered worse when the data consist of less time points, when the reference profiles are more congruent, and when the data contains more error (see Fig. 6).

5. Application

Clinical trials commonly study the effect of a particular biomedical or behavioral intervention. Whereas this effect may be different for different symptoms or persons [36], closer investigations at either symptom- or person level are still the exception. In this paper we will focus on a single mental disorder, namely major depressive disorder, which is highly prevalent and puts a strong burden on society [37]. Especially for the treatment of MDD, insights into differential symptom and/or person effects could greatly advance the field, seeing that a number of

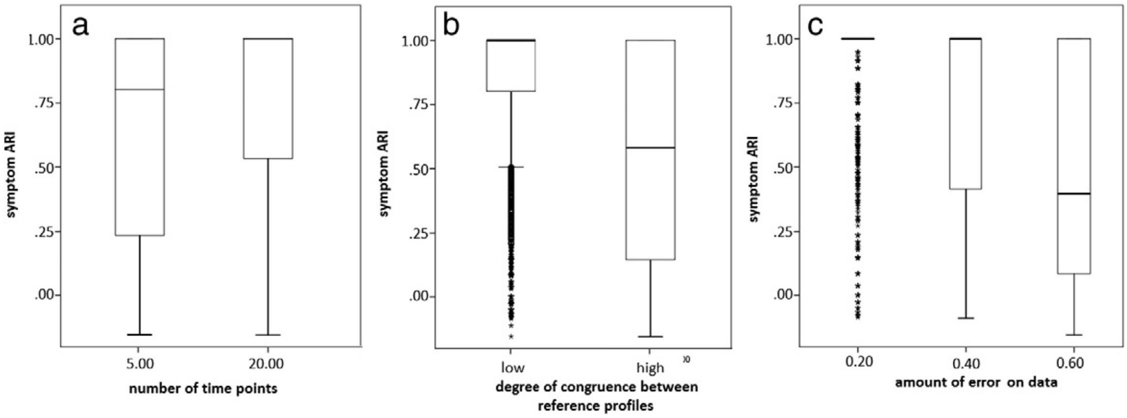

Fig. 6. Boxplots of the symptom ARI as a function of (a) the number of time points, (b) the congruence between reference profiles and (c) the amount of error on the data.

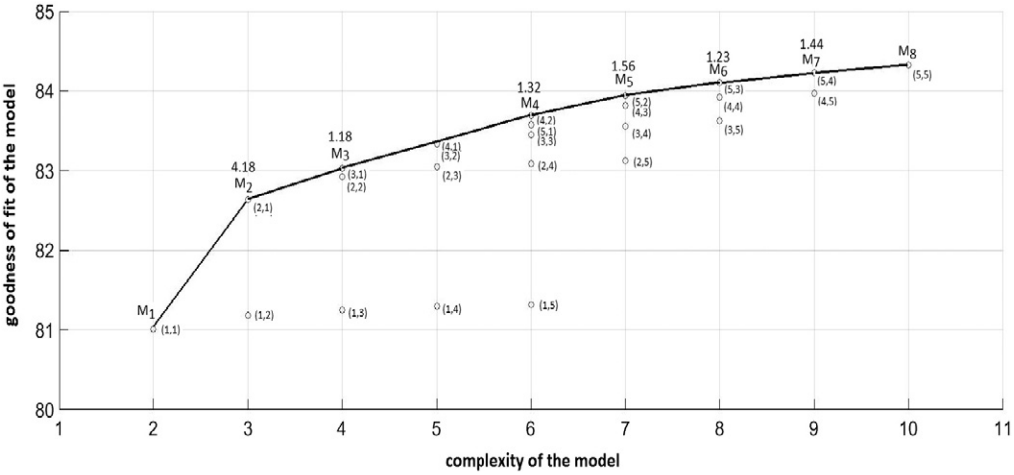

Fig. 7. Graphical representation of the complexity of the obtained 2M-KSC solutions for the STAR*D data (i.e., sum of person and symptom clusters) versus their goodness of fit (i.e., percentage of sum of squares explained). The full line represents the higher boundary of the convex hull. The *st* values of the eight hull solutions are displayed above this line.

large meta-analyses have shown that antidepressants only slightly outperform placebos [38,39,40,41]. Indeed, these meta-analyses focus on the main effect of antidepressants, while the size of this effect is moderated by person characteristics as well as symptom features (e.g., [42,43]). Therefore, with this application, we aim to examine whether persons can be divided into a number of person clusters and whether symptoms can be divided into a number of symptom clusters, where combinations of a person and a symptom cluster are associated with distinct symptom severity time profiles. Taken together, we hope to uncover symptom and person differences in response to the intervention.

To this end, we analyze data set version 3.0 from the NIH-supported STAR*D study [3,4][4]. In this study, a total of 1240 MDD patients received the selective serotonin reuptake inhibitor (SSRI), citalopram for at least five subsequent clinical visits[5] (week 0, 2, 4, 6, and 9 of the treatment). However, within this group of 1240 patients a lot of data were missing not at random[6]. To adequately handle this we would have to implement multiple imputation methods. This extension falls outside the scope of this manuscript, as this application is merely a proof-of-principle study. Consequently, we listwise deleted persons with missing scores. This way, we end up analyzing the symptom time profiles (five time points) of 169 MDD patients. To monitor treatment during the STAR*D study, clinicians rated the Quick Inventory of Depressive Symptoms (QIDS; [44]), containing the following depression symptoms: (1) problems falling asleep, (2) waking up during the night, (3) waking up too early, (4) sleeping too much, (5) feeling sad, (6) decreased or increased appetite, (7) decreased or increased weight, (8) lack of concentration and difficulties with decision making, (9) negative view of myself, (10) thoughts of death or suicide, (11) lack of general interest, (12) low energy level, (13) feeling slowed down, and (14) feeling restless. At entry and exit, some additional characteristics were evaluated: overall MDD symptom severity, functional outcome, quality of life, side effects, patient satisfaction and utilization and cost, assessing the number of times a healthcare provider or Emergency Room was visited or the number of times a hospital admission was required for both mental health and other medical reasons.

To find an optimal two-mode KSC model for this dataset, we applied the CHull procedure described in Section 3.3 followed by split-half stability analyses. More specifically, we analyzed our data with the 2M-KSC algorithm with both K, the number of person clusters, and C, the number of symptom clusters, varying from one to five. For each value of K and C, we used 500 random initializations and 1 rational initialization[7] of both the person and symptom partitions. Fig. 7 displays the percentage of sum of squares accounted for by the different obtained (K, C) solutions, as a function of their complexity $K + C$. Furthermore, the scree test ratios *st* for the 8 models on the upper boundary of the convex hull of this plot, are shown. Next, we examined the stability of the four hull models with the highest scree ratios: the (2,1) model, the (4,2) model, the (5,2) model, and the (5,4) model.

For this purpose, we randomly split the person mode of the data into two halves and analyzed both resulting data halves with 2M-KSC (and repeated this procedure ten times). To decide whether or not a solution should be retained for further inspection, we examined the stability of the person clustering, by comparing the person clusterings resulting from the analyses of the two halves with the person clustering resulting from the analysis of the full data set. To deal with the permutational freedom of the person and symptom clusterings, we evaluated all possible permutations and retained the one that maximized the mean Tucker congruence value between the reference profiles from the analyses of the halves and those from the analysis of the full data set. Next, we counted the number of persons for whom the cluster membership switched from the analysis of the full data to that of the data halves. The frequencies of data splits for which these numbers were observed are displayed in Fig. 8(a), for each of the four considered model complexities. This Figure reveals that the person clustering of the (2,1) model is rather stable, with the number of persons switching clusters varying from 1 to 10. This is clearly not the case for the (4,2), (5,2), and (5,4) solutions. Based on these results the models with complexity (4,2), (5,2), and (5,4) will no longer be considered, and the (2,1) model should be retained.

However, from an illustration point of view, the (2,1) solution is not ideal, since it contains a single symptom cluster only. Therefore, we will discuss the (2,2) solution, since it fits only slightly worse than the (3,1) solution (which is the hull solution of complexity four), since it has a stable person clustering (see Fig. 8(b); only 3 to 11 persons switched person cluster) which moreover closely resembles that of the (2,1) solution (only 10 persons are in a different cluster), and since it tells apart two types of symptoms.

[4] This manuscript reflects the views of the authors and may not reflect the opinions or views of the STAR*D Study Investigators or the NIH.

[5] This amount of time points was chosen to have enough points to reliably cluster the data.

[6] This is due to several reasons (e.g., treatment discontinuation due to decrease or increase of symptoms, or switching treatment).

[7] Note, that in computing the single rational start in this application, profiles with a constant severity score of zero were treated as profiles with a constant positive severity score (>0) in order to be able to scale the profile to a SSQ value of one.

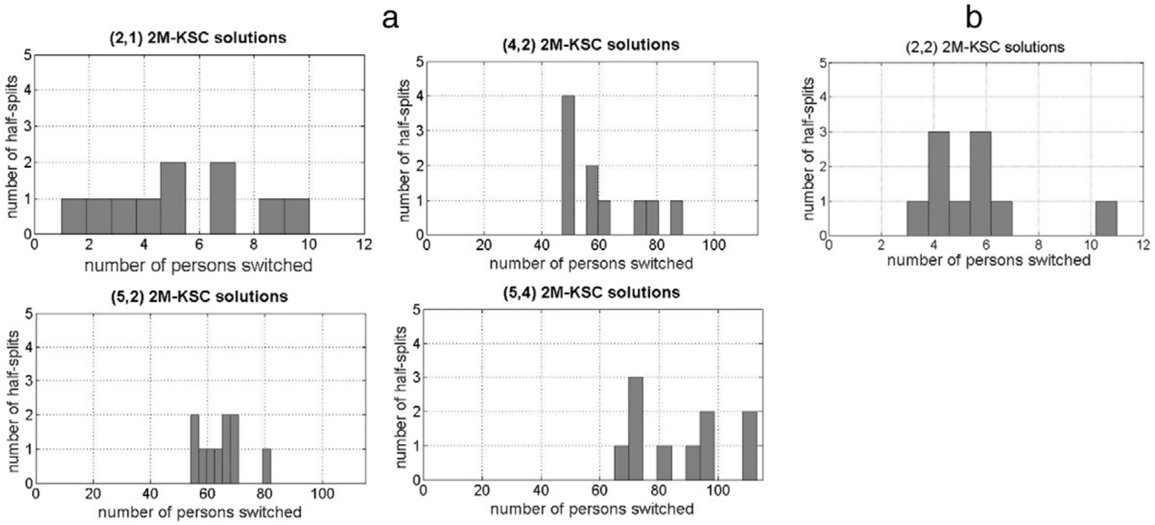

Fig. 8. Histograms of the number of persons switching clusters in the split-half stability analyses: (a) histograms for the (2,1), (4,2), (5,2), and (5,4) solutions and (b) histogram for the (2,2) solutions.

In the (2,2) solution, out of the 169 MDD patients, 80 were assigned to the first person cluster and 89 to the second one. Regarding the clustering of symptoms, we find that 10 symptoms are assigned to the first symptom cluster and four to the second symptom cluster. As the second symptom cluster comprises problems with falling asleep, waking up during the night, sleeping too much, and feeling restless, we call this the 'agitation hampering sleep' symptom cluster, while the first symptom cluster will be referred to as the general symptoms cluster. It is instructive to look at the stability of this symptom clustering in the split-half stability analyses. In Fig. 9, the gray bars reflect in how many analyses of data halves the symptom was assigned to the first symptom cluster, while the dashed bars indicate how often the symptom was assigned to the second symptom cluster. This Figure suggests that the symptom clustering is rather stable with the first cluster always containing the symptoms 'waking up during the night' and 'feeling restless', mostly completed with the symptoms 'sleeping too much', 'problems falling asleep' and 'low energy level'. The second cluster always contains the symptoms 'feeling sad', 'increased or decreased appetite', 'increased or decreased weight', 'negative view of myself', 'thoughts on death or suicide', and 'feeling slowed down', mostly completed with the symptoms 'lack of general interest', 'waking up too early', and 'lack of concentration'.

The four obtained reference profiles are displayed in Fig. 10. From this Figure we conclude that the severity of both symptom clusters decreases from time point 1 to time point 5 for patients in the first person cluster. However, clients in the second person cluster show severity time profiles that stagnate, meaning that the intervention does not really change the symptom severity. Therefore the first person cluster will be referred to as the better response cluster, while the second one will be labeled the worse response cluster. Within the better response cluster we also see differences in symptom time profiles for agitation hampering sleep and general symptoms, namely that the severity decrease is higher for general symptoms than for agitation hampering sleep symptoms, implying that agitation symptoms are more persistent for these patients.

We used the characteristics measured at exit of the treatment to validate the patient clustering, by means of point-biserial correlation coefficients. For five characteristics the absolute value of this correlation

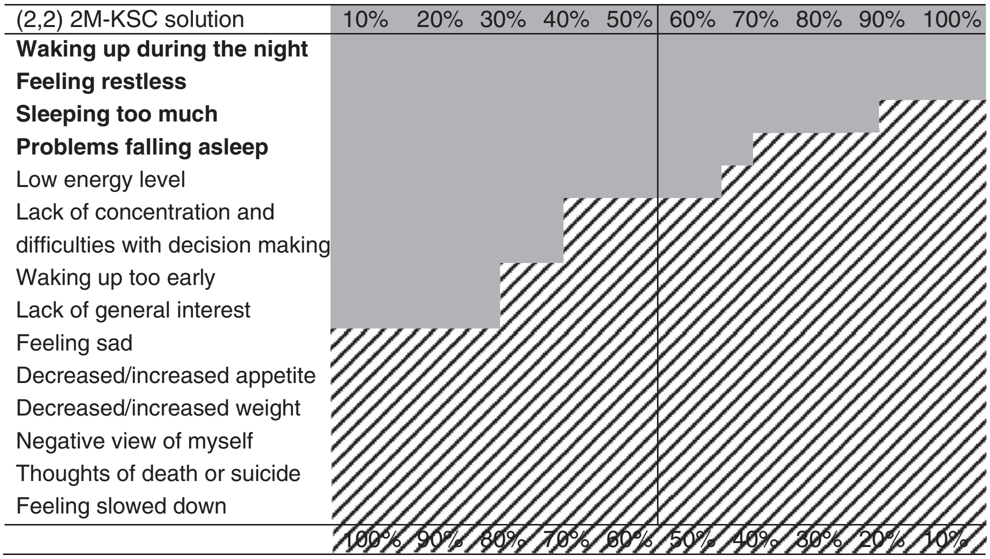

Fig. 9. Stability of the symptom clustering in the split-half stability analyses of the (2,2) solutions. Gray bars reflect in how many data halves the symptom was assigned to the first symptom cluster, while the dashed bars indicate how often the symptom was assigned to the second symptom cluster.

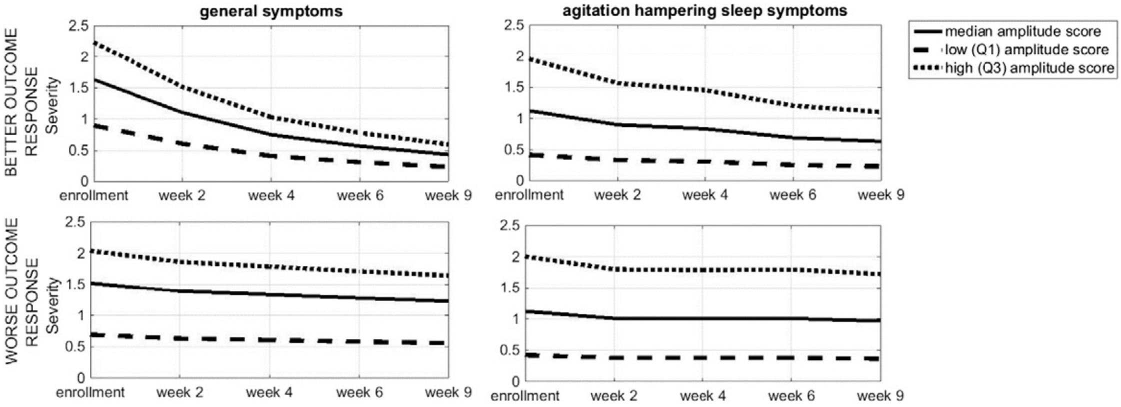

Fig. 10. Obtained reference profiles for the MDD application.

was stronger than .50; boxplots of these variables are shown in Fig. 11. We conclude that, at treatment exit, the patients from the better response cluster showed a better outcome in terms of MDD symptoms (Hamilton Rating Scale for Depression, $r_{pb} = .55$; Inventory of Depressive Symptomatology – Clinician Rated, $r_{pb} = .55$; Inventory of Depressive Symptomatology – Self Rated, $r_{pb} = .50$), functioning (Short Form Health Survey – Mental, $r_{pb} = -.50$) and quality of life (Quality of Life Enjoyment and Satisfaction, $r_{pb} = -.50$) than those in the worse response cluster.

Overall, while the small number of patients examined here cannot be considered representative of all participants in the STAR *D study, or of depressed patients in general, we draw three conclusions from our proof-of-principle study that may be worth investigating in follow-up studies and larger samples. First, patients in the first person cluster respond better to the citalopram treatment as they show a stronger relative decrease in symptom severity. Second the effect of citalopram is stronger on general depression symptoms than on agitation hampering sleep symptoms. Third patients in the first cluster show a higher functional outcome and quality of at treatment exit than patients in the second person cluster.

6. Discussion

In this discussion, we will first discuss some other application areas, where 2M-KSC might be useful. Next, we go deeper into the model characteristics and compare 2M-KSC to other biclustering methods at the phenotype level (see Introduction).

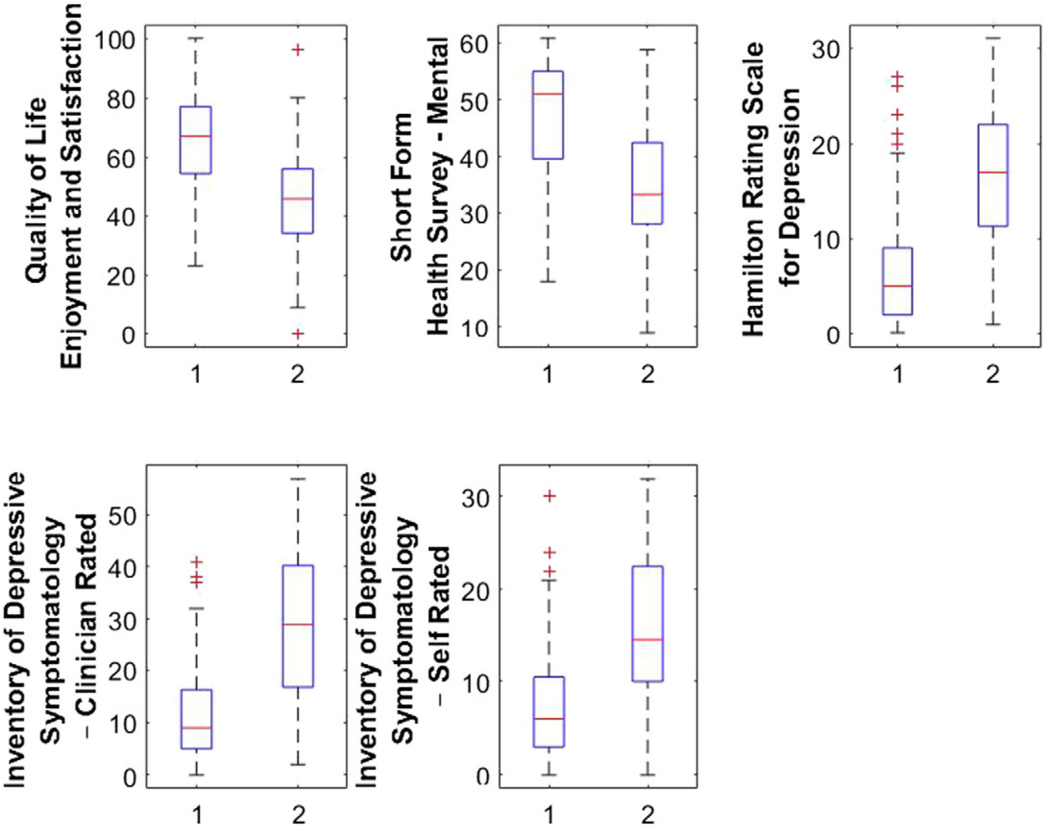

Fig. 11. Boxplots of five patient characteristics as a function of person cluster, with 1 indicating the better response cluster and 2 the worse response cluster.

6.1. Other domains of application

Intervention studies are only one of the possible fields of application of 2M-KSC. In psychology and special education, prospective studies have gained much attention in the last years, in which multiple variables are tracked across time for multiple persons. For example, many psychological studies track mental health of persons (e.g., students) across time in multiple-wave longitudinal designs, measuring several aspects of psychological well-being at each time point (e.g., [45]). Similarly, in special education, researchers aim to gain insight into developmental problems such as mathematical learning difficulties by exploring the associations between arithmetic strategy development, numerical magnitude processing, working memory and phonological processing across time (e.g., [46]).

2M-KSC could also be of use in systems biology and metabolomics where the presence of an abundance of biological variables (e.g., biomolecules, metabolites) in biological samples (e.g., urine, blood) is tracked across time, often under different experimental conditions (e.g., [47]). 2M-KSC would allow to unravel whether or not the partitioning of the samples and variables is respectively related to the different experimental conditions and biological processes, while simultaneously revealing differential effects of the experimental conditions on these processes.

Moreover in environmental studies, one often measures multiple pollution parameters (e.g., water quality, air quality, presence of different chemicals) across time at different geographical sites. The question then rises whether biclusters of sites and pollution parameters can be found which are characterized by distinct evolution profiles of pollution across time. This is another question that could be tackled by 2M-KSC.

Finally, in signal processing, databases are available on the amount of emails a set of persons send to each other across time (e.g., the famous ENRON email corpus; [48]). Applying 2M-KSC would allow to inspect whether groups of senders and receivers can be discerned for which email correspondence for instance peaks at different points in time.

6.2. Similarities and differences between 2M-KSC and other methods at the phenotype level

As stated in the Introduction, interest in clustering the multivariate time profiles of different persons has strongly increased. In this paper, we are particularly interested in biclustering methods that can simultaneously cluster both persons and variables, based on the phenotype of the associated time profiles. Thus, although a much larger literature on biclustering methods exists (e.g. [49,50,51,52,53]), we will focus on methods that are developed to find biclusters in a longitudinal three way setting (e.g., [48,54,55,56,57,58]). In the next paragraphs we review the differences and similarities between 2M-KSC and these other phenotype level methods, making use of three distinguishing characteristics.

A first distinguishing characteristic pertains to the nature of the clustering of the persons and the variables, and the resulting biclustering. Regarding the clustering of the persons and the variables, this clustering can either be exclusive (each person/variable belongs to one cluster at maximum) or overlapping (each person/variable can belong to multiple clusters) on the one hand, and partial (not all persons/variables are clustered) or complete (all persons/variables are clustered) on the other hand, and may be different for persons and variables. If at least one of both clusterings is exclusive, the biclusters are exclusive as well (panels a and b of Fig. 12). In case both the person and variable clustering are overlapping, the associated biclustering can be exclusive (panel c of Fig. 12) or overlapping (panel d of Fig. 12). Whereas 2M-KSC implies an exclusive and complete person and variable clustering (panel a of Fig. 12), many of the alternative methods induce overlapping and partial clusterings (e.g., [48,54,55,56,57,58]) because most of them were built to study gene expression. Indeed, it makes sense that single genes can be associated with multiple biological functions or processes [59] while others do no play a role in the processes under study.

The second characteristic pertains to the type of biclusters, as biclusters can either be homogeneous, heterogeneous on the persons or heterogeneous on the variables. When a bicluster is homogeneous this means that all time profiles within the same bicluster are modeled by exactly the same reference profile, as is the cased in 2M-KSC or in both the coclustering method of Papalexakis, Sidiropoulos and Bro [48] and the tensor factorization methods of Li, Ye, Wu, and Ng [55], and Zhang, Wang, Ashby, Chen and Huang [58]. When a bicluster is heterogeneous on the persons (resp. variables), each person (resp. variable) in the bicluster has its own reference profile, but this reference profile stays the same for all variables (resp. persons) within the bicluster. Gene expression focused methods use in most cases heterogeneous biclusters. For instance, Jiang et al. [54] look for biclusters of genes and conditions (i.e., modules) in which each gene displays the same time profile for all conditions in the module, but the profiles of the separate genes may differ, while Polanski et al. [56] rather allow for heterogeneity across conditions. Furthermore, Supper et al. [57] induce homogeneous as well as heterogeneous biclusters.

The third and last characteristic pertains to whether or not the method is built on a fit measure that indicates how well the total data set is reconstructed. 2M-KSC models the total data set by simultaneously partitioning all rows and columns into mutually exclusive clusters. To this end a loss function is used that indicates how much of the variability in the observed data is captured by the model. In contrast, most, but not all (e.g., [48]), alternative methods sequentially extract biclusters, without considering how well they fit the complete data (e.g. [54,56,57],).

Table 2 summarizes these similarities and differences between the different methods. Note that which method is most appropriate depends on the particular data set at hand and the associated research questions. For instance, an alternative research question about the STAR*D data could be which subset of symptoms are similarly

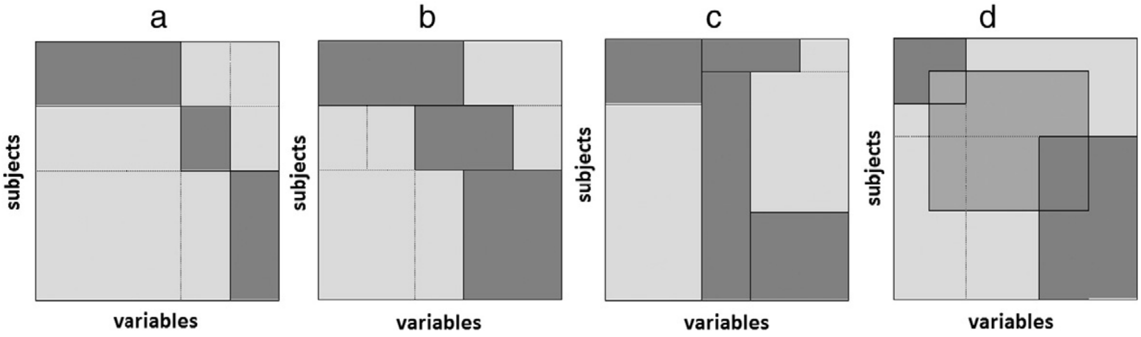

Fig. 12. Possible assignment patterns: (a) exclusive person and variable clustering, (b) exclusive person and overlapping variable clustering, (c) overlapping person and variable clustering with exclusive biclustering and, (d) overlapping person and variable clustering with overlapping biclustering; for both partial (dark gray coloring) and complete clustering (gray coloring and shading).

Table 2
Overview of the similarities and differences between 2M-KSC and related methods.

Method	Nature of clustering	Type of biclusters	Reconstruction total data set
2M-KSC (2016)	Exclusive + complete	Homogeneous	Yes
Papalexakis, Sidiropoulos, & Bro (2013) [48]	Overlapping + partial	Homogeneous	Yes
Jiang, Pei, Ramanathan, Tang, & Zhang (2004) [54]	Overlapping + partial	Heterogeneous	No
Li, Ye, Wu, & Ng (2012) [55]	Overlapping + partial	Homogeneous	Yes
Polanski, Rhodes, Hill, Zhang, Jenkins, Kiddle, et al. (2014) [56]	Overlapping + partial	Heterogeneous	No
Supper, Strauch, Wanke, Harter, & Zell (2007) [57]	Overlapping + partial	Homogeneous + heterogeneous	No
Zhang, Wang, Ashby, Chen, & Huang (2012) [58]	Overlapping + partial	Homogeneous	Yes

influenced by the citalopram medication for a subset of MDD patients (where the exact form of this influence may differ from patient to patient within the subset, implying heterogeneity), without making any claims about the similarity of these symptoms for the other patients (implying that the clustering may be partial and that no full reconstruction of the data set is needed). In that case we could opt to use either the method of Jiang et al. [54] or the method of Polanski et al. [56].

6.3. Conclusion

We introduced 2M-KSC to study how the shape of multivariate time profiles varies as a function of the persons and variables under study. To this aim, persons and variables are partitioned simultaneously, relating each combination of a person and variable cluster, and thus each bicluster, to a single reference profile. This reference profile reflects the prototypical shape of the profiles in the bicluster, discarding amplitude scaling differences. Such differences are modeled by means of amplitude scores.

Acknowledgments

This research was supported by Grant GOA/15/003 from the Research Fund of the University of Leuven and by a research grant from the Fund for Scientific Research-Flanders (FWO, Project No. G.0582.14 awarded to Eva Ceulemans, Peter Kuppens and Francis Tuerlinckx). The research leading to the results reported in this paper was supported in part by the Interuniversity Attraction Poles program financed by the Belgian government (IAP/P7/06).

Data used in the preparation of this article were obtained from the limited access datasets distributed from the NIH-supported “Sequenced Treatment Alternatives to Relieve Depression” (STAR*D). STAR*D focused on non-psychotic major depressive disorder in adults seen in outpatient settings. The primary purpose of this research study was to determine which treatments work best if the first treatment with medication does not produce an acceptable response. The study was supported by NIMH Contract # N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is NCT00021528.

References

[1] B.L. Andersen, W.B. Farrar, D.M. Golden-Kreutz, R. Glaser, C.F. Emery, T.R. Crespin, et al., Psychological, behavioral, and immune changes after a psychological intervention: a clinical trial, J. Clin. Oncol. 22 (17) (2004) 3570–3580.

[2] K. Kajander, K. Hatakka, T. Poussa, M. Färkkilä, R. Korpela, A probiotic mixture alleviates symptoms in irritable bowel syndrome patients: a controlled 6-month intervention, Aliment. Pharmacol. Ther. 22 (5) (2005) 387–394.

[3] M. Fava, A.J. Rush, M.H. Trivedi, A.A. Nierenberg, M.E. Thase, H.A. Sackeim, et al., Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study, Psychiatr. Clin. N. Am. 26 (2) (2003) 457–494.

[4] A.J. Rush, M. Fava, S.R. Wisniewski, P.W. Lavori, M.H. Trivedi, H.A. Sackeim, et al., Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design, Control. Clin. Trials 25 (1) (2004) 119–142.

[5] T.W. Liao, Clustering of time series data — a survey, Pattern Recogn. 38 (11) (2005) 1857–1874.

[6] A.K. Smilde, J.A. Westerhuis, H.C.J. Hoefsloot, S. Bijlsma, C.M. Rubingh, D.J. Vis, et al., Dynamic metabolomic data analysis: a tutorial review, Metabolomics 6 (1) (2010) 3–17.

[7] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 2006.

[8] G. Claeskens, B.W. Silverman, L. Slaets, A multiresolution approach to time warping achieved by a Bayesian prior–posterior transfer fitting strategy, J. R. Stat. Soc. Ser. B (Stat Methodol.) 72 (5) (2010) 673–694.

[9] B.L. Jones, D.S. Nagin, Advances in group-based trajectory modeling and an SAS procedure for estimating them, Sociol. Methods Res. 35 (2007) 542–571.

[10] G.J. Palardy, J.K. Vermunt, Multilevel growth mixture models for classifying groups, J. Educ. Behav. Stat. 35 (2010) 532–565.

[11] M. Wang, T.E. Bodner, Growth mixture modeling identifying and predicting unobserved subpopulations with longitudinal data, Organ. Res. Methods 10 (2007) 635–656.

[12] N.A. Heard, C.C. Holmes, D.A. Stephens, D.J. Hand, G. Dimopoulos, Bayesian coclustering of anopheles gene expression time series: study of immune defense response to multiple experimental challenges, Proc. Natl. Acad. Sci. U. S. A. 102 (47) (2005) 16939–16944.

[13] S.P. Ellner, B.E. Kendall, S.N. Wood, E. McCauley, C.J. Briggs, Inferring mechanism from time-series data: delay-differential equations, Physica D Nonlin. Phenom. 110 (3) (1997) 182–194.

[14] R. Yoshida, S. Imoto, T. Higuchi, Estimating time-dependent gene networks from time series microarray data by dynamic linear models with Markov switching, Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE, IEEE 2005, pp. 289–298.

[15] J. Yang, J. Leskovec, Patterns of temporal variation in online media, Proceedings of the Fourth ACM International Conference on Web Search and Data Mining 2011, pp. 177–186.

[16] J. Heylen, P. Verduyn, I. Van Mechelen, E. Ceulemans, Variability in anger intensity profiles: structure and predictive basis, Cogn. Emotion 29 (1) (2015) 168–177.

[17] P.M. Kroonenberg, J. De Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms, Psychometrika 45 (1980) 69–97.

[18] H.A.L. Kiers, Hierarchical relations among three-way methods, Psychometrika 56 (1991) 449–470.

[19] E. Ceulemans, I. Van Mechelen, Tucker2 hierarchical classes analysis, Psychometrika 69 (2004) 375–399.

[20] E. Ceulemans, I. Van Mechelen, Hierarchical classes models for three-way three-mode binary data: interrelations and model selection, Psychometrika 70 (2005) 461–480.

[21] J. Heylen, I. Van Mechelen, P. Verduyn, E. Ceulemans, KSC-N: clustering of hierarchical time profile data, Psychometrika (2016) (in press).

[22] E. Ceulemans, I. Van Mechelen, I. Leenen, The local minima problem in hierarchical classes analysis: an evaluation of a simulated annealing algorithm and various multistart procedures, Psychometrika 72 (2007) 377–391.

[23] D. Steinley, Local optima in K-means clustering: what you don't know may hurt you, Psychol. Methods 8 (2003) 294–304.

[24] J. Schepers, I. Van Mechelen, E. Ceulemans, Three-mode partitioning, Comput. Stat. Data Anal. 51 (2006) 1623–1642.

[25] E. Ceulemans, H.A.L. Kiers, Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method, Br. J. Math. Stat. Psychol. 59 (2006) 133–150.

[26] E. Ceulemans, H.A.L. Kiers, Discriminating between strong and weak structures in three-mode principal component analysis, Br. J. Math. Stat. Psychol. 62 (2009) 601–620.

[27] T.F. Wilderjans, E. Ceulemans, K. Meers, CHull: a generic convex hull based model selection method, Behav. Res. Methods 45 (2013) 1–15.

[28] R.B. Cattell, The scree test for the number of factors, Multivar. Behav. Res. 1 (2) (1966) 245–276.

[29] J. Schepers, E. Ceulemans, I. Van Mechelen, Selecting among multi-mode partitioning models of different complexities: a comparison of four model selection criteria, J. Classif. 25 (2008) 67–85.

[30] M.J. Brusco, J.D. Cradit, A variable selection heuristic for K-means clustering, Psychometrika 66 (2001) 249–270.

[31] M.J. Brusco, J.D. Cradit, ConPar: a method for identifying groups of concordant subject proximity matrices for subsequent multidimensional scaling analyses, J. Math. Psychol. 49 (2005) 142–154.

[32] S. Hands, B. Everitt, A monte carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques, Multivar. Behav. Res. 22 (1987) 235–243.

[33] G.W. Milligan, S.C. Soon, L.M. Sokol, The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure, IEEE Trans. Pattern Anal. Mach. Intell. 5 (1983) 40–47.

[34] K. De Roover, E. Ceulemans, M.E. Timmerman, K. Vansteelandt, J. Stouten, P. Onghena, Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data, Psychol. Methods 17 (2012) 100–119.
[35] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218.
[36] E.I. Fried, R.M. Nesse, Depression sum-scores don't add up: why analyzing specific depression symptoms is essential, BMC Med. 13 (72) (2015) 1–11.
[37] World Health Organization, Integrating Mental Health Into Primary Care: A Global Perspective, World Health Organization, 2008.
[38] A. Khan, W.A. Brown, Antidepressants versus placebo in major depression : an overview, World Psychiatry 14 (2015) 294–300.
[39] A. Khan, S. Khan, W.A. Brown, Are placebo controls necessary to test new antidepressants and anxiolytics? Int. J. Neuropsychopharmacol. 5 (2002) 193–197.
[40] I. Kirsch, B.J. Deacon, T.B. Huedo-Medina, A. Scoboria, T.J. Moore, B.T. Johnson, Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration, PLoS Med. 5 (2) (2008), e45.
[41] H.E. Pigott, A.M. Leventhal, G.S. Alter, J.J. Boren, Efficacy and effectiveness of antidepressants: current status of research, Psychother. Psychosom. 79 (5) (2010) 267–279.
[42] F. Hieronymus, J.F. Emilsson, S. Nilsson, E. Eriksson, Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression, Mol. Psychiatry (2015) 1–8.
[43] N. Kaymaz, J. van Os, A.J.M. Loonen, W.A. Nolen, Evidence that patients with single versus recurrent depressive episodes are differentially sensitive to treatment discontinuation: a meta-analysis of placebo-controlled randomized trials, J. Clin. Psychiatry 69 (2008) 1423–1436.
[44] A.J. Rush, M.H. Trivedi, H.M. Ibrahim, T.J. Carmody, B. Arnow, D.N. Klein, et al., The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression, Biol. Psychiatry 54 (5) (2003) 573–583.
[45] M.L. Pe, A. Brose, I.H. Gotlib, P. Kuppens, Affective Updating Ability and Stressful Events Interact to Prospectively Predict Increases in Depressive Symptoms Over Time, 2015 Emotion.
[46] K. Vanbinst, P. Ghesquière, B. De Smedt, Arithmetic strategy development and its domain-specific and domain-general cognitive correlates: a longitudinal study in children with persistent mathematical learning difficulties, Res. Dev. Disabil. 35 (2014) 3001–3013.
[47] M.E. Timmerman, H.C.J. Hoefsloot, A.K. Smilde, E. Ceulemans, Scaling in ANOVA-simultaneous component analysis, Metabolomics 11 (2015) 1265–1276.
[48] E.E. Papalexakis, N. Sidiropoulos, R. Bro, From k-means to higher-way co-clustering: multilinear decomposition with sparse latent factors, IEEE Trans. Signal Process. 61 (2013) 493–506.
[49] S.C. Madeira, A.L. Oliveira, A linear time biclustering algorithm for time series gene expression data, Proceedings of Fifth Workshop on Algorithms in Bioinformatics 2005, pp. 39–52.
[50] S.C. Madeira, A.L. Oliveira, A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series, Algorithms Mol. Biol. 4 (1) (2009) 8.
[51] S.C. Madeira, M.C. Teixeira, I. Sa-Correia, A.L. Oliveira, Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm, Comput. Biol. Bioinform. 7 (1) (2010) 153–165.
[52] J. Meng, Y. Huang, Biclustering of time series microarray data, Methods Mol. Biol. 802 (2012) 87–100.
[53] I. Van Mechelen, H.-H. Bock, P. De Boeck, Two-mode clustering methods: a structured overview, Stat. Methods Med. Res. 13 (2004) 363–394.
[54] D. Jiang, J. Pei, M. Ramanathan, C. Tang, A. Zhang, Mining coherent gene clusters from gene-sample-time microarray data, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2004, pp. 430–439.
[55] X. Li, Y. Ye, Q. Wu, M.K. Ng, Multifactv: finding modules from higher-order gene expression profiles with time dimension, 2012 IEEE International Conference on Bioinformatics and Biomedicine 2012, pp. 1–6.
[56] K. Polanski, J. Rhodes, C. Hill, P. Zhang, D.J. Jenkins, S.J. Kiddle, et al., Wigwams: identifying gene modules co-regulated across multiple biological conditions, Bioinformatics 30 (7) (2014) 962–970.
[57] J. Supper, M. Strauch, D. Wanke, K. Harter, A. Zell, EDISA: extracting biclusters from multiple time-series of gene expression profiles, BMC Bioinf. 8 (1) (2007) 334.
[58] S. Zhang, K. Wang, C. Ashby, B. Chen, X. Huang, A unified adaptive co-identification framework for high-D expression data, Pattern Recognition in Bioinformatics 2012, pp. 59–70.
[59] A. Tanay, R. Sharan, R. Shamir, Discovering statistically significant biclusters in gene expression data, Bioinformatics 18 (Suppl. 1) (2002) S136–S144.