Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 537–549
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A unifying model involving a categorical and/or dimensional reduction for multimode data

Iven Van Mechelen*, Jan Schepers

SymBioSys, Katholieke Universiteit Leuven, Belgium

Available online 12 March 2007

Abstract

A unifying model is presented that implies a categorical and/or dimensional reduction of one or several modes of a multiway data set. The model encompasses a broad range of (existing as well as to be developed) discrete, continuous, as well as hybrid discrete–continuous reduction models as special cases, which all imply a decomposition of the reconstructed data on the basis of quantifications of the different data modes and a linking array. An analysis of the objective or loss function associated with the model leads to two generic algorithmic strategies, the possibilities and limitations of which are the object of a subsequent discussion.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Clustering; Dimension reduction; Multiway; Decomposition; Unifying model

1. Introduction

Data that imply one or more sets of entities (or modes) with a large number of elements (experimental units, variables, time points, others) imply a major challenge for the data analyst. This is even more the case if the data pertain to more than two modes, that is, if they are multiway multimode in nature. The complexity of the information as present in such data may be tremendous. In order to grasp it in a proper way, the data analyst may wish to subject one or more of the data modes to a (simultaneous) reduction. Reduction is here to be understood either in a categorical sense, in that the elements of the reduced mode are grouped into a small number of clusters (which may be overlapping or not, and which may cover the full mode or not), or in a dimensional sense, in that the elements of the reduced mode are represented as points in a lowdimensional space. A simultaneous reduction further can be purely categorical, that is, categorical for all reduced modes, purely dimensional, that is, dimensional for all reduced modes, or hybrid, that is, categorical for some of the reduced modes and dimensional for the other ones. Purely categorical reduction models can be amply found in the clustering domain, examples including one-mode partitioning models (such as *k*-means type models and all kinds of one-mode mixture models, e.g., McLachlan and Chang, 2004), two-mode clustering (or biclustering) models (such as two-mode hierarchical and additive clustering models, Furnas, 1980; Gaul and Schader, 1996, and two-mode hierarchical classes models, De Boeck and Rosenberg, 1988; Van Mechelen et al., 1995), as well as their multimode generalizations (e.g., Ceulemans and Van Mechelen, 2005; Eckes and Orlik, 1994). Pure dimension reduction models can be amply found in the domain of component and factor analysis, examples including the standard two-mode principal component model and its multimode generalizations (such as PARAFAC/CANDECOMP and the

* Corresponding author. Tel.: +32 16 326131; fax: +32 16 325993.
E-mail address: Iven.VanMechelen@psy.kuleuven.be (I. Van Mechelen).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.03.001

family of N-mode Tucker models, e.g., Kroonenberg, 1983). Examples of hybrid models include various projection pursuit type clustering methods (e.g., Bock, 1987; Vichi and Kiers, 2001), cluster differences scaling (Heiser and Groenen, 1997), and cluster unfolding (De Soete and Heiser, 1993).

The family of categorical and dimensional reduction models for multimode data clearly is very large in number. Moreover, it is also fairly heterogeneous, both in terms of the mathematical structures implied by the different models and by the principles and methods used in the associated data analysis. In the present paper, we will contribute to a clarification of this situation by introducing a unifying model that encompasses a broad range of (existing as well as to be developed) discrete, continuous and hybrid reduction models as special cases. The to be proposed unifying model considerably extends the already very broad CANDCLUS and MUMCLUS models as proposed by Carroll and Chaturvedi (1995), with this extension including a much broader family of decomposition functions other than (generalized) Cartesian products, room for various types of modeling constraints, and room for a possible addition of distributional assumptions. An analysis of the objective or loss function associated with the unifying model will further lead to two generic algorithmic strategies, the possibilities and limitations of which are the object of a subsequent discussion.

The remainder of this paper is organized as follows: In Section 2 we will introduce the type of data under study, along with a few associated concepts. In Section 3 we will introduce our unifying reduction model. The associated objective or loss function will be dealt with in Section 4 and the algorithmics in Section 5. Section 6 will present a general discussion.

2. Data

Data arrays can have different conceptual structures. In order to typify the various cases, Carroll and Arabie (1980) have introduced some terminology (which in turn relies on work by Tucker, 1964). To use this terminology, a data set is conceived as a mapping $\mathbf{D}$ from a Cartesian product $S = S_1 \times S_2 \times \cdots \times S_N$ of N sets $S_1, \ldots, S_N$ to some (typically univariate) domain Y: for any N-tuple $(s_1, s_2, \ldots, s_N)$ with $s_1 \in S_1, \ldots, s_N \in S_N$ a value $\mathbf{D}(s_1, s_2, \ldots, s_N)$ from Y is recorded. The total number N of constituent (possibly identical) sets of S is called the number of *ways* in the data, whereas the number of *distinct* sets in S is called the number of *modes*. In the present paper we will limit ourselves to data arrays for which all ways pertain to different sets of entities, yielding N-way N-mode data. We will further also assume that the domain of the data mapping $\mathbf{D}$ coincides with the full Cartesian product S, herewith excluding data structures for which prespecified parts of S are structurally missing, such as nested data structures and data stemming from between-subject designs.

A second important data characteristic pertains to which data entries are comparable, that is, to the level of conditionality of the data. In principle, data entry comparability can be limited, for instance, to entries pertaining to the same variable (whereas values pertaining to different variables are not comparable—a case also referred to as exemplifying a lack of commensurability). Comparability or conditionality is not totally unrelated to a possible preprocessing of the data, in that several authors implicitly assume that by means of a suitable preprocessing (e.g., by subjecting each variable to a z-transformation), full data array conditionality may be restored; the latter, however, is a far from trivial assumption, which may be the subject of a difficult debate. Anyhow, in the remainder of this paper, we will assume full data array conditionality (possibly upon a suitable preprocessing of the data).

A third important aspect pertains to the goal that is associated with the data. Within this paper, we aim at a data analysis that includes a reduction of one or more of the data modes, with this reduction being optimal in some sense. A general definition of optimality in this regard could refer to minimizing the loss of information as implied by the reduction. However, at this point different information aspects as included in the data can be distinguished. For example, N-mode data can be interpreted as the values of a single criterion variable Y as a function of N nominal predictor variables, each of the predictor variables corresponding to one of the data modes; in that case, one may consider a categorical reduction of each of the data modes in terms of a partition, the resulting N-mode partition being optimal in that the averaged values of the criterion variable per N-mode partition class imply a minimal loss of information on the amount of interaction between the predictor variables in the prediction of Y, as present in the original data array (Bock, 1979). As an alternative information aspect, one may wish the reduction to be such that it allows for an *optimal reconstruction* of the full data array (optimality to be formalized in, e.g., a least L_p or a maximum likelihood sense). In this paper we will focus on a data reconstruction type of optimality.

3. Model

Assume a real-valued $(I_1 \times \cdots \times I_n \times \cdots \times I_N)$ N-way N-mode data array $\mathbf{D}$ (i.e., a mapping $\mathbf{D} : S_1 \times S_2 \times \cdots \times S_N \rightarrow \mathbb{R}$, with $\#S_n = I_n$) with entries $d_{i_1 \ldots i_n \ldots i_N}$. The unifying reduction model we propose for $\mathbf{D}$ includes a deterministic heart and optional additional stochastic assumptions. We will now successively introduce both. Subsequently, we will discuss how various existing reduction models show up as special cases of the unifying model.

3.1. Deterministic heart

The deterministic heart of the unifying model includes a decomposition based on the following three constituents: (1) N component matrices, one for each mode $(A^1, \ldots, A^n, \ldots, A^N)$, with A^n being of size $I_n \times P_n$, and $(P_1, \ldots, P_n, \ldots, P_N)$ denoting the complexity or rank of the decomposition; (2) a $P_1 \times \cdots \times P_n \times \cdots \times P_N$ core array $\mathbf{W}$; (3) a decomposition function f. The decomposition then reads as follows:

$$\mathbf{D} = f(A^1, \ldots, A^N, \mathbf{W}) + \mathbf{E}, \tag{1}$$

with $\mathbf{E}$ denoting an $(I_1 \times \cdots \times I_n \times \cdots \times I_N)$ array with error entries.

- The component matrices A^n can take values in (a) {0,1} (categorical or crisp clustering reduction of mode n), (b) [0,1] (fuzzy clustering reduction of mode n), or (c) $\mathbb{R}$ (dimensional reduction of mode n). A special case shows up if mode n is left unreduced; in that case A^n equals an identity matrix.
- The core array $\mathbf{W}$ can take values in (a) {0,1}, (b) $\mathbb{R}$ or $\mathbb{R}^+$.
- The decomposition function f should be such that for any mode $n = 1, \ldots, N$ it holds that

$$f(A^1, \ldots, A^N, \mathbf{W})_{i_1 \ldots i_n \ldots i_N} \quad \text{depends only on the } i_n\text{th row of } A^n, \tag{2}$$

meaning that $f(A^1, \ldots, A^N, \mathbf{W})_{i_1 \ldots i_n \ldots i_N}$ depends only on the pattern of component scores for the i_nth element of S_n (and not on the patterns for the other elements of S_n). Apart from this constraint, the function f can take a broad range of possible forms. Below we illustrate this by means of four clearly distinct classes of functions that are very often used in practice:

(a) *A generalized Cartesian product*:

$$f(A^1, \ldots, A^N, \mathbf{W})_{i_1 \ldots i_n \ldots i_N} = \sum_{p_1=1}^{P_1} \cdots \sum_{p_n=1}^{P_n} \cdots \sum_{p_N=1}^{P_N} a^1_{i_1 p_1} \ldots a^n_{i_n p_n} \ldots a^N_{i_N p_N} w_{p_1 \ldots p_n \ldots p_N}. \tag{3}$$

(b) *A generalized Cartesian product in a Max, times structure*:

$$f(A^1, \ldots, A^N, \mathbf{W})_{i_1 \ldots i_n \ldots i_N} = \operatorname*{Max}_{p_1=1}^{P_1} \ldots \operatorname*{Max}_{p_n=1}^{P_n} \ldots \operatorname*{Max}_{p_N=1}^{P_N} a^1_{i_1 p_1} \ldots a^n_{i_n p_n} \ldots a^N_{i_N p_N} w_{p_1 \ldots p_n \ldots p_N}. \tag{4}$$

(c) *A weighted Minkowski distance*; *for example*, *assuming* $N = 3$ *and* $A^3 \geqslant 0$, $\mathbf{W} \geqslant 0$:

$$f(A^1, A^2, A^3, \mathbf{W})_{i_1 i_2 i_3} = \left[\sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} \sum_{p_3=1}^{P_3} a^3_{i_3 p_3} w_{p_1 p_2 p_3} |a^1_{i_1 p_1} - a^2_{i_2 p_2}|^k \right]^{1/k}. \tag{5}$$

(d) *A dominance operator*; *for example*, *assuming* $N = 2$, $\mathbf{W} \geqslant 0$:

$$f(A^1, A^2, \mathbf{W})_{i_1 i_2} = \operatorname*{Min}_{p_1=1}^{P_1} \operatorname*{Min}_{p_2=1}^{P_2} [w_{p_1 p_2} I_{(a^1_{i_1 p_1} \geqslant a^2_{i_2 p_2})}], \tag{6}$$

with I denoting an indicator variable.

Apart from the above specifications, the component matrices and core array can either be left unconstrained or can be subject to a broad range of constraints. As to the latter, linking up with a framework developed by Ceulemans

et al. (2004) (see also Bro, 1998; Kiers and Smilde, 1998), one may draw a distinction between value and structure constraints. Value constraints imply that all or some of the values of a component matrix (resp. the core array) are fixed on prespecified values. Structure constraints imply that some structural aspect of a component matrix (resp. the core array) is preserved, without fixing prespecified entries.

An example of value constraints on a component matrix is that, in a confirmatory study, some or all values are kept fixed on values as obtained from the analysis of some previous data set. With regard to the core array, if $P_1 = \cdots = P_n = \cdots = P_N$ a possible set of value constraints is implied by putting $\mathbf{W}$ equal to a superidentity array (i.e., $w_{p_1 \ldots p_n \ldots p_N} = 1$ iff $p_1 = \cdots = p_n = \cdots = p_N$, otherwise $w_{p_1 \ldots p_n \ldots p_N} = 0$) or to an array with superdiagonal structure.

Examples of structure constraints on a binary component matrix include the cases in which that matrix is constrained to be a partition matrix, a membership matrix of a nested clustering, or a Guttman scale, the latter denoting a matrix the rows of which are totally ordered in terms of the natural pointwise order defined on vectors ($\mathbf{a} \leqslant \mathbf{a}'$ iff $\forall i : a_i \leqslant a'_i$) (Guttman, 1944). Examples of structure constraints on a core array include the case in which the core has to take a small, prespecified number of nonzero entries, and the case in which the core can be subjected to some low rank decomposition.

3.2. Optional stochastic assumptions

In addition to the deterministic core of the model, optional additional assumptions can be made with regard to the distribution of the entries of the component matrices A^n and of the error array $\mathbf{E}$. In this regard, one may consider a two-stage sampling model:

The first stage pertains to the sampling of the rows of the A^n-matrices. The rows of these matrices can either be treated as fixed, or as random samples according to some distribution. Assuming binary entries, possible distributions in the latter case include a discrete uniform and a multinomial distribution.

The second stage pertains to the sampling of the data entries given the entries of the A^n-matrices. One possibility in this regard could be

$$d_{i_1 \ldots i_n \ldots i_N} \sim N(f(A^1, \ldots, A^N, \mathbf{W})_{i_1 \ldots i_n \ldots i_N} \sigma^2) \quad \text{and independent,}$$

or,

$$e_{i_1 \ldots i_n \ldots i_N} \overset{iid}{\sim} N(0, \sigma^2).$$

3.3. Substantive justification of the model and special cases

In terms of a substantive justification, decomposition rule (1) may be looked at as a formal tool to reveal the structure of each of the data modes as well as their interrelation. As an additional option, in several cases the rule (as well as the component matrices and core array of the model) could also formalize a substantive mechanism underlying the data. Below, we illustrate with special cases subsumed by the unifying model. In this regard we focus on one specific model for each of the decomposition functions (3)–(6). For each model we will specify the key model equation along with a hypothetical example to sketch the model's potential substantive relevance. Subsequently, we will make clear how the specific model in question belongs to a broader class of models subsumed by the unifying model.

Generalized Cartesian product (3): We start with the special case obtained for $N = 3$ and with the component matrices A^n as well as the core array $\mathbf{W}$ taking values in $\mathbb{R}$. This yields

$$f(A^1, A^2, A^3, \mathbf{W})_{i_1 i_2 i_3} = \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} \sum_{p_3=1}^{P_3} a^1_{i_1 p_1} a^2_{i_2 p_2} a^3_{i_3 p_3} w_{p_1 p_2 p_3}, \tag{7}$$

which is the model equation of the Tucker3 model. Assuming spectroscopic data from chemometrics on emissions at I_1 different wavelengths (S_1) from a set of I_2 chemical compounds or batches (S_2) at I_3 different time points during a chemical reaction (S_3), model (7), along with $\mathbf{W}$ being constrained to be a superidentity array, may reveal the spectra of constituents of the compounds (as formalized by A^1), the compositions of the compounds (as formalized by A^2), and the impact of the chemical reaction on the constituents (as formalized by A^3). In this context the Sum-operator as

included in (7) further appears to be the most natural operator to link the spectra of the constituents to the spectra of the compounds as included in the data.

One may note that if in (7) the three component matrices are constrained to be binary partition matrices of S_1, S_2, and S_3, this yields the three-mode partitioning model as studied by Kiers (2004), Rocci and Vichi (2005), and Schepers et al. (2006). More in general, decomposition function (3) with the component matrices A^n taking binary values or values in $\mathbb{R}$, and without making any stochastic assumptions yields the (deterministic) MUMCLUS model as mentioned by Carroll and Chaturvedi (1995), whereas the same model with **W** being constrained to be superdiagonal yields the (deterministic) CANDCLUS model as studied more in detail by the same authors.

Interestingly, in this case (as well as in all other cases subsumed by the unifying model) the deterministic core of the model mentioned above may also be extended with distributional assumptions, yielding stochastic counterparts of the model in question. For example, (7) with the three component matrices being constrained to binary partition matrices can be extended with the assumption that the rows of these matrices are randomly sampled from a multinomial distribution as well as with the assumption of *iid* normally distributed error terms, yielding a three-mode mixture model; this model could be considered a stochastic counterpart of one particular instance of Carroll and Chaturvedi's MUMCLUS model.

Generalized Cartesian product in a Max, times structure (4): We start here with the case in which $N = 2$ and the component matrices A^n as well as the core matrix **W** are binary. In that case the core matrix can, without loss of generality, be put equal to the identity matrix. This yields

$$f(A^1, A^2)_{i_1 i_2} = \operatorname*{Max}_{p_1=1}^{P_1} a^1_{i_1 p_1} a^2_{i_2 p_1} = \bigoplus_{p_1=1}^{P_1} a^1_{i_1 p_1} a^2_{i_2 p_1}, \tag{8}$$

with $\oplus$ denoting Boolean sum. This is the key equation of the hierarchical classes model as proposed by De Boeck and Rosenberg (1988). Assuming psychometric data on success/failure (1/0) of a set of persons (S_1) on a set of items or cognitive tasks (S_2), model (8) may reveal the solution strategies that each of the persons masters or not (as formalized by A^1), and through which the tasks can be solved or not (as formalized by A^2). In this context the Max-operator as included in (8) formalizes the idea that a person will succeed in a task iff there is at least one strategy that the person masters and through which the problem can be solved.

As regards the general case with arbitrary values $N \geqslant 2$, one may note that if in (4) all component matrices A^n are constrained to be binary and the core array **W** is constrained to be binary or positive real-valued, this yields the general family of hierarchical classes models as studied extensively by Leenen et al. (1999), Ceulemans et al. (2003) and Van Mechelen et al. (in press). Interestingly, a distinctive feature of this family is that the binary component matrices A^n are constrained to represent if–then type relations between the elements of the corresponding mode (e.g., success on task a implies success on task b).

Weighted Minkowski distance: We start here from (5) and assume $k = 2$, $N = 3$, component matrices A^n taking values in $\mathbb{R}$, $A^3 \geqslant 0$, and **W** superidentity. This yields

$$f(A^1, A^2, A^3)_{i_1 i_2 i_3} = \left[\sum_{p_1=1}^{P_1} a^3_{i_3 p_1} (a^1_{i_1 p_1} - a^2_{i_2 p_1})^2 \right]^{1/2}, \tag{9}$$

which is the equation of three-way metric unfolding (see, e.g., DeSarbo and Carroll, 1985). Assuming data from marketing research in which a set of I_1 brands (S_1) is judged on a set of I_2 characteristics (S_2) by a set of I_3 consumers (S_3), this model may reveal the coordinates of brands and characteristics (as formalized by A^1 and A^2) in some joint perceptual space, the dimensions of which are of differential importance for distinct consumers (as formalized by A^3).

More in general, through decomposition functions involving a weighted Minkowski distance, the unifying model (1) subsumes a broad class of scaling and unfolding models as special cases. All models of this class share the representation of a number of reduced data modes in some lowdimensional metric space.

The dominance operator: We start from (6) and assume $N = 2$, component matrices A^n taking values in $\mathbb{R}$, and **W**=identity matrix. This yields

$$f(A^1, A^2)_{i_1 i_2} = \operatorname*{Min}_{p_1=1}^{P_1} [I_{(a^1_{i_1 p_1} \geqslant a^2_{i_2 p_1})}], \tag{10}$$

which is the equation of conjunctive nonmetric factor analysis (Coombs, 1964). Assuming again psychometric data on success/failure (1/0) of a set of persons (S_1) on a set of items or cognitive tasks (S_2), model (10) may reveal P_1

underlying cognitive competence dimensions, on each of which each person takes some position (the ability levels as formalized by A^1), whereas the same holds for the tasks (the difficulty levels as formalized by A^2). Decomposition rule (10) then formalizes the natural dominance idea that a person will succeed in a task iff for all underlying cognitive dimensions the ability level of that person exceeds the difficulty level posed by the task in question.

Other models based on a dominance operator include the disjunctive model of nonmetric factor analysis and the ORDCLAS models as proposed by Leenen et al. (2001). Further application areas in which dominance relations show up in quite a natural way include the modeling of choice behavior.

3.4. The unifying model as a powerful tool for model development

As amply illustrated in the previous section, our unifying model subsumes a broad range of existing model families as special cases. Beyond this, however, it may also act as a powerful and versatile tool for the development of novel models. In this section, we will illustrate this with one specific example.

As a guiding application for the to be proposed novel model, one may think of bioinformatics data pertaining to the mRNA expression level of a set of genes (S_1) with regard to a set of tissues or experimental conditions (S_2). As in this context relations between genes, between conditions, as well as between genes and conditions are all three of biological importance, a simultaneous representation of all three relations by an unfolding model may be most useful (Van Deun et al., 2007). Such a model can be based on the following decomposition rule:

$$f(A^1, A^2)_{i_1 i_2} = \left[\sum_{p_1=1}^{P_1} (a^1_{i_1 p_1} - a^2_{i_2 p_1})^2 \right]^{1/2}. \tag{11}$$

It implies a joint representation of genes and conditions as points in a lowdimensional space, with the amount of expression of gene i_1 in condition i_2 being (inversely) related to the Euclidean distance between gene point i_1 and condition point i_2. In many microarray gene expression applications, however, both the number of genes and the number of conditions may be prohibitively large for this type of representation. In such situations, one could consider a novel extension of the unfolding model (11), which implies a partition of both genes and conditions into a limited number of partition classes as well as a joint representation of these classes as points in a lowdimensional Euclidean space. This can be formalized as follows:

$$f(A^1, A^2, \mathbf{W})_{i_1 i_2} = \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} a^1_{i_1 p_1} a^2_{i_2 p_2} w_{p_1 p_2} \tag{12}$$

subject to the constraints that: (a) A^1 and A^2 are binary partition matrices, and (b) that

$$w_{p_1 p_2} = \left[\sum_{p_1^*=1}^{P_1^*} (a^{*1}_{p_1 p_1^*} - a^{*2}_{p_2 p_1^*})^2 \right]^{1/2}, \tag{13}$$

with A^{*1} and A^{*2} denoting two additional $P_1 \times P_1^*$ and $P_2 \times P_1^*$ component matrices. All this means that the gene expression data are subject to a two-mode partitioning model (12) (see, e.g., Van Mechelen et al., 2004), subsumed by the unifying model as a special case of (3), with the core matrix $\mathbf{W}$ of this model in its turn being subject to a Euclidean-distance based decomposition constraint (13), subsumed by the unifying model as a special case of (5). This means that the proposed novel model implies a double, recursion-like use of the unifying modeling framework. Obviously, on the basis of a similar recursion-like strategy, quite a novel class of two-mode clustering methods could be devised.

4. Criterion to be optimized in the data analysis

In the deterministic case, the objective or loss function l to be minimized in the data analysis will typically be of the least L_p type,

$$l(A^1, \ldots, A^N, \mathbf{W}) = \sum_{i_1,\ldots,i_n,\ldots,i_N} |d_{i_1\ldots i_n\ldots i_N} - (f(A^1, \ldots, A^N, \mathbf{W})_{i_1\ldots i_n\ldots i_N})|^p, \tag{14}$$

with, in case $p = 2$:

$$l(A^1, \ldots, A^N, \mathbf{W}) = \sum_{i_1,\ldots,i_n,\ldots,i_N} [d_{i_1\ldots i_n\ldots i_N} - (f(A^1, \ldots, A^N, \mathbf{W})_{i_1\ldots i_n\ldots i_N})]^2. \tag{15}$$

In the stochastic case, the objective function to be maximized will be the likelihood. In this regard, it may be useful to note that for models of real-valued data maximizing the likelihood is equivalent to minimizing the least squares loss function (15), provided that the following three conditions hold:

(1) The rows of the A^n-matrices are treated as fixed or as random according to a uniform distribution.
(2) If the rows of the A^n-matrices are treated as random, the entries of these matrices are treated as to be estimated parameters (generalized classification likelihood approach).
(3) $e_{i_1\ldots i_n\ldots i_N} \overset{iid}{\sim} N(0, \sigma^2)$ (which includes the assumptions of homoscedasticity and independence of the errors).

5. Algorithmics

5.1. Two propositions

A first proposition focuses on the core array $\mathbf{W}$:

Proposition 1. *If* $\mathbf{W}$ *is real-valued and unconstrained,* f *equals a generalized Cartesian product* (3), *and the loss function equals* (15), *then the conditionally optimal* $\mathbf{W}$, *given component matrices* $(A^1, \ldots, A^n, \ldots, A^N)$, *can be expressed as a closed form function of the component matrices,* $g(A^1, \ldots, A^n, \ldots, A^N)$.

Proof. The conditionally optimal $\mathbf{W}$ can be considered a set of regression weights in the prediction of the vectorized data on the basis of $P_1 \times \cdots \times P_n \times \cdots \times P_N$ predictor vectors that result from the multiplication of the columns of the component matrices. For example, in case $N = 3$ and denoting by $A^n_{.p_n}$ the p_nth column of A^n, the $P_1 \times P_2 \times P_3$ predictor vectors equal:

$$A^3_{.p_3} \otimes A^2_{.p_2} \otimes A^1_{.p_1},$$

with $\otimes$ denoting the Kronecker product (Kiers, 2000). From standard least squares regression theory it then follows that $\mathbf{W}$ can be obtained as

$$\mathbf{W}_{A^1} = [(A^{1'}A^1)^{-1}A^{1'}]\mathbf{D}_{A^1}\{[A^3(A^{3'}A^3)^{-1}] \otimes [A^2(A^{2'}A^2)^{-1}]\},$$

with $\mathbf{W}_{A^1}$ and $\mathbf{D}_{A^1}$ denoting the A^1-mode matricized core and data arrays (with dimensions $P_1 \times P_2P_3$ and $I_1 \times I_2I_3$, respectively) (Kiers, 2000). □

In all cases in which a closed form solution $g(A^1, \ldots, A^n, \ldots, A^N)$ exists for the conditionally optimal $\mathbf{W}$, given component matrices $(A^1, \ldots, A^n, \ldots, A^N)$, this solution can be plugged in loss function (14), yielding a new loss function, $\tilde{l}(A^1, \ldots, A^n, \ldots, A^N)$, with

$$\tilde{l}(A^1, \ldots, A^N) = l(A^1, \ldots, A^N, g(A^1, \ldots, A^n, \ldots, A^N)). \tag{16}$$

For this loss function the following useful proposition holds:

Proposition 2. *Minimizing* $l(A^1, \ldots, A^N, \mathbf{W})$ *over* $A^1, \ldots, A^N, \mathbf{W}$ *is equivalent to minimizing* $\tilde{l}(A^1, \ldots, A^N)$ *over* $A^1, \ldots, A^N$.

Proof. This proposition immediately follows from the fact that $(A^1, \ldots, A^N, \mathbf{W})$ being a global optimum of l implies that $\mathbf{W}$ is conditionally optimal given $(A^1, \ldots, A^N)$. □

A third proposition involves the concept of separability:

Definition. Given parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\varphi}$, a loss function $l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\varphi})$ is called *separable* with regard to subsets of parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K$ if, conditional upon any value of $\boldsymbol{\varphi}$, the conditional optimization of $l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\varphi})$ with respect to $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K$ can be done by minimizing for each $k = 1, \ldots, K$ a partial loss function $l_k(\boldsymbol{\theta}_k, \boldsymbol{\varphi})$ with respect to $\boldsymbol{\theta}_k$, with $l_k\ (\boldsymbol{\theta}_k, \boldsymbol{\varphi})$ not depending on $\boldsymbol{\theta}_{k'}\ \forall k' \neq k$.

The proposition below makes use of this definition, with the I_n rows of one of the A^n-matrices in the role of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K$.

Proposition 3. *If the component matrices* $(A^1, \ldots, A^n, \ldots, A^N)$ *are not subject to constraints that simultaneously involve multiple rows*, *then for each n loss function* (14) *is separable with respect to the rows of* A^n.

Proof. From (14), it follows that:

$$
\begin{aligned}
&l(A^1, \ldots, A^N, \mathbf{W}) \\
&\quad = \sum_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_N} |d_{i_1\ldots i_{n-1},1,i_{n+1}\ldots i_N} - (f(A^1, \ldots, A^N, \mathbf{W})_{i_1\ldots i_{n-1},1,i_{n+1}\ldots i_N})|^p \\
&\qquad + \cdots + \sum_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_N} |d_{i_1\ldots i_{n-1},i_n,i_{n+1}\ldots i_N} - (f(A^1, \ldots, A^N, \mathbf{W})_{i_1\ldots i_{n-1},i_n,i_{n+1}\ldots i_N})|^p \\
&\qquad + \cdots + \sum_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_N} |d_{i_1\ldots i_{n-1},I_n,i_{n+1}\ldots i_N} - (f(A^1, \ldots, A^N, \mathbf{W})_{i_1\ldots i_{n-1},I_n,i_{n+1}\ldots i_N})|^p.
\end{aligned}
$$

For the decomposition function f it further holds that for any mode $n = 1, \ldots, N f(A^1, \ldots, A^N, \mathbf{W})_{i_1\ldots i_n\ldots i_N}$ only depends on the i_nth row of A^n. If the component matrices of the model under study are not subject to constraints that simultaneously involve multiple rows of A^n, then it further follows that the conditional optimization of $l(A^1, \ldots, A^N, \mathbf{W})$ with respect to the I_n rows of A^n can be done by minimizing separately for each value of $i_n = 1, \ldots, I_n$:

$$
\sum_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_N} |d_{i_1\ldots i_n\ldots i_N} - (f(A^1, \ldots, A^N, \mathbf{W})_{i_1\ldots i_{n-1},i_n,i_{n+1}\ldots i_N})|^p,
$$

with respect to the i_nth row of A^n. □

Interestingly, Proposition 2 holds for all decomposition functions f that satisfy constraint (2). Moreover, it is easy to show that the proposition can be extended to the likelihood-based loss function for stochastic instances of the unifying model, provided an independence assumption on the error terms $e_{i_1\ldots i_n\ldots i_N}$. Note, however, that the proposition does *not* hold for loss function (16).

To clarify the concept of "constraints that simultaneously involve multiple rows", we note, as a first illustrative example, that a constraint such as "every row of a component matrix should contain one nonzero value" is a constraint that does not simultaneously involve multiple rows. As a second illustrative example, a constraint upon a binary component matrix such as "this matrix should have the structure of a nested clustering" does simultaneously involve multiple rows.

At this point, a warning may further be needed. In practice, the condition of Proposition 2 will in many cases not be met, strictly speaking. More in particular, when optimizing loss function (14) over sets of component matrices

$(A^1, \ldots, A^n, \ldots, A^N)$ and a core array $\mathbf{W}$ with some prespecified complexity or rank $(P_1, \ldots, P_n, \ldots, P_N)$, one may wish that the resulting solution is of full rank, that is, that there should not exist a second set of component matrices $(A^{*1}, \ldots, A^{*n}, \ldots, A^{*N})$ and a core array $\mathbf{W}^*$ with rank $(P_1^*, \ldots, P_n^*, \ldots, P_N^*)$ and such that:

(a) $f(A^1, \ldots, A^N, \mathbf{W}) = f(A^{*1}, \ldots, A^{*N}, \mathbf{W}^*)$.
(b) $\forall n : P_n^* \leqslant P_n$ with at least one inequality being strict.

This condition is, for instance, not met, in case of decomposition function (3), if some A^n is not of full column rank. Clearly, being of full column rank (which includes absence of zero columns as an important special case) is a condition that simultaneously involves multiple rows. In algorithmic practice, however, situations in which the condition of Proposition 2 is not met because of a rank constraint only, do not necessarily imply a severe problem. Indeed, during an iterative algorithmic process one may temporarily leave the rank constraint aside (herewith restoring the condition of Proposition 2 and therefore the resulting separability of the loss function). One may then simply trust or prove that, when the iterative algorithm arrives at a rank-deficient solution during the course of the algorithmic process, it will return to a full-rank solution at subsequent iterations; as an alternative, an additional algorithmic step may be devised in order to force such a return; still another option could be to make use of a multistart strategy in which runs that end up in rank-deficient solutions are naturally discarded.

We finally note that Proposition 2 generalizes earlier findings on separability of Carroll and Chaturvedi (1995) to a much broader class of models. Otherwise, Carroll and Chaturvedi did not take the issue of rank deficiency into account; as such, the warning above also implies a qualification for their separability claims.

5.2. Two generic algorithmic strategies

Various algorithmic strategies could be considered to estimate instances of the unifying model, including, for example, several kinds of local search methods such as simulated annealing. In this subsection, we will focus on two generic algorithmic strategies that may be used to minimize loss functions (14) and (16), respectively. Both algorithms are of an alternating least L_p type. This means that, after an initialization, they consist of an iterated conditional updating until convergence is reached. In each step the value of the loss function will remain constant or will decrease. In principle, there is only a guarantee for convergence on the level of the loss function and not on the level of the resulting model solution. In general, a conditional updating step can pertain to an algorithmic (re)*calculation* of a subset of the model parameters that necessarily yields a conditionally optimal estimate for them, or to a heuristic (re)*estimation* of the subset in question that may not yield a conditionally optimal solution. Irrespective of whether conditional updates are conditionally optimal or not, upon eventual convergence, the end result of the iterated process always may be a locally optimal solution only.

Strategy 1: Optimization of $l(A^1, \ldots, A^N, \mathbf{W})$

This strategy consists of the following three steps:

1. Initialization of $(A^1, \ldots, A^n, \ldots, A^N)$.
2. Conditional calculation or estimation of $\mathbf{W}$.
3. Subsequent conditional recalculation or re-estimation of $A^1, \ldots, A^n, \ldots, A^N, \mathbf{W}$.

Step 3 is to be repeated until the loss function does not change any more across two update cycles of the full model.

To fill out this strategy, choices have to be made on three different levels:

- The type of initialization procedure, which may be random, rational, or somewhere in between.
- The procedure for the conditional (re)estimation of $\mathbf{W}$ if a closed form solution is not available (which implies that the conditions of Proposition 1 are not met).
- The procedure for the conditional re-estimation of $A^1, \ldots, A^n, \ldots, A^N$ if a closed form solution is not available.

With regard to the third choice to be made, typically, if l is separable, it will be possible to calculate conditionally optimal updates for each of the rows of A^n, for example, by making use of a suitable quasi-enumerative procedure if A^n is binary, or by making use of a regression if A^n is real-valued and f is a generalized Cartesian product function.

Strategy 2: Optimization of $\tilde{l}(A^1, \ldots, A^N)$
This strategy consists of the following three steps:

1. Initialization of $(A^1, \ldots, A^n, \ldots, A^{N-1})$.
2. Conditional estimation of A^N.
3. Subsequent conditional re-estimation of $A^1, \ldots, A^n, \ldots, A^N$.

Step 3 is to be repeated until the loss function does not change any more across two update cycles of the full model.
To fill out this strategy, choices have to be made on two different levels:

- The type of initialization procedure, which may be random, rational, or somewhere in between.
- The procedure for the conditional re-estimation of $A^1, \ldots, A^n, \ldots, A^N$.

In general, in the second strategy the conditional (re-)estimations imply a major algorithmic challenge. A possible group of heuristic procedures may be obtained by acting as if $\tilde{l}$ is separable, that is, by going for a rowwise updating of A^n. One should, however, take care with such an updating procedure, in that the result may, amongst other things, depend on the order of the rows at hand.

6. Discussion

In this paper we introduced a novel unifying model for multimode data based on two key components: (a) the elements of each of the modes involved in the data are reduced to either points in a lowdimensional space or to elements of a limited set of (possibly overlapping) clusters, and (b) the connection between the dimensions and clusters to which each of the data modes are reduced is captured by a linking array. The reduction for each of the modes is further such that the coordinates or cluster memberships for any element of any data mode are "private" in the sense that reconstructed data entries pertaining to that element depend on the coordinates or cluster memberships of that element only, and not on those of any other elements of the same mode.

The proposed unifying model clearly subsumes a very broad range of existing reduction models—discrete, continuous, as well as hybrid; deterministic as well as stochastic. Advantages of the unification include: (a) the establishment of a formal framework in which model interrelations (e.g., relations of nestedness) can be mathematically specified (see, e.g., Ceulemans and Van Mechelen, 2005); (b) the facilitation of transfers of insights that have been obtained for particular subclasses of the unifying model to other subclasses (as an example, one may think of the implementation of constraints that have been devised within the context of particular subclasses of the unifying model—see, e.g., Bro, 1998; Ceulemans et al., 2004—to quite different subclasses); (c) the algorithmic result implied by Proposition 3, which holds for a very broad range of instances subsumed by the generic model.

The unifying model further also implies a powerful and versatile tool for the development of new modeling instances that may be particularly useful within various substantive contexts. In the present paper we illustrated this with the formulation of a novel type of a two-mode partitioning/unfolding model that appears most relevant for the modeling of large-sized data sets as these show up in challenging research domains such as bioinformatics. Obviously, new to be developed instances of the unifying model could be deterministic as well as stochastic, with stochastic extensions of existing deterministic models as a special option.

In the present paper we gave a generic description of the unifying model and two possible associated algorithmic strategies. However, many important aspects of the unifying model have not been addressed. Below we mention some of them more in detail:

- Identifiability: For each instance of the unifying model, identifiability is to be studied. In general, it should be clear that, without imposing additional constraints, many instances of the unifying model are not identified. As examples, one may think of a broad range of instances relying on a decomposition function that involves a generalized Cartesian product (viz., decomposition function (3)) or a weighted Euclidean distance (viz., decomposition function (5) with $k = 2$), which are subject to rotational freedom. Obviously, nonidentifiability of this and similar types does not imply the need to discard the models in question; yet, it implies challenges with regard to the specification of limitations of model inferences, to the search of additional constraints under which

identifiability could be restored, and to the search of necessary and/or sufficient (data- or model-dependent) identifiability conditions.

- Model selection: The rich modeling flexibility that is implied by the unified modeling framework of course obviously yields a major challenge on the level of model selection, both with regard to the deterministic heart of the model and with regard to possible additional stochastic assumptions. With regard to the deterministic heart, model selection turns out to be a highly multifaceted issue, with facets referring to: (a) which modes are to be reduced and which not, what the nature of the reduction should be for each to be reduced mode (discrete or continuous), as well as what should be the extent of the reduction (number of dimensions or clusters to be used), (b) the decomposition function to be used, and (c) the choice of possible value or structure constraints on the component matrices and/or the core array. Obviously, it is impossible to devise a single model selection strategy to deal with all these choice aspects simultaneously. Choices in general can be based on quite different bases, including: a priori preferences of researchers and research domains (e.g., for dimensional rather than discrete reductions or vice versa); a priori theoretical interests (e.g., in if–then type relations, the representation of which may necessitate a decomposition with a Max- or Min-operator—see Van Mechelen et al., 1995) or a priori theoretical hypotheses with regard to the decomposition mechanism underlying the data (see, e.g., the different substantive mechanisms as included in the examples of Section 3.3, justifying quite different decomposition functions); a priori data characteristics, such as data size, which may be of direct importance in guiding the choice of the extent of data reduction (see, e.g., the extensive two-step reduction as implied by the newly proposed two-mode partitioning/unfolding model to fit very large-sized microarray gene expression data sets); and data characteristics that can be included in data-driven model selection strategies (e.g., in order to decide between different model complexities or between different decomposition functions). Model selection further also is an issue on the level of the optional stochastic assumptions that may be made in any instance of the unifying model. Choice aspects here include the decision as to whether making such assumptions or not and, if yes, what their nature should be. As to the latter, sequential model building strategies could also be considered that start from a (possibly overly simple but in practice very often used) homoscedastic model with *iid* residuals from some standard (e.g., Gaussian) distribution, and that subsequently move to more complex assumptions if suitable model checks point at a need for those.
- Choice of loss function: In case of deterministic instances of the unifying model with an L_p-norm based loss function, the value of p needs to be specified. Possible bases for this choice may include: (a) substantive concerns (with, e.g., certain applications with rating-valued data naturally leading to an L_1-norm); (b) robustness concerns (which may lead to a preference for values of p smaller than 2); (c) computational ease: for example, the desire for a closed-form expression for the conditionally optimal $\mathbf{W}$, given component matrices $(A^1, \ldots, A^n, \ldots, A^N)$, may lead to the choice of $p = 2$ (on the basis of Proposition 1). As to the latter, one should note that $p = 2$ is in general not a necessary condition for the existence of such a solution (counterexamples including, e.g., multimode partitioning models, for which conditional medians act as a closed form estimator of $\mathbf{W}$ in case of an L_1-norm based loss function). Otherwise, generic singly necessary and jointly sufficient conditions for the existence of a closed-form conditional estimator of $\mathbf{W}$ seem to be out of reach.
- Choice of an algorithmic strategy: We proposed two generic and broadly applicable strategies to solve the optimization problem associated with instances of the unifying model. The second strategy requires a closed-form expression for the conditionally optimal $\mathbf{W}$. If such an expression is available, a choice has to be made between the two generic algorithmic strategies. At first sight, the first strategy may look preferable, especially because the implied loss function looks easier to deal with, and because, in view of Proposition 3, it may often be separable (possibly upon neglection of the rank deficiency issue), which may considerably simplify the conditional updating process. At second sight, however, a cautionary note may be needed. Indeed, in recent simulation studies, Schepers et al. (2006) compared several instantiations of the two generic strategies as briefly outlined above, for the particular case of the three-mode partitioning model (i.e., for the case with $N = 3$, A^1, A^2, A^3 partition matrices, $\mathbf{W}$ taking values in $\mathbb{R}$, and f being a generalized Cartesian product). In a first simulation study, they found that instantiations of the first strategy slightly outperformed instantiations of the second one. In subsequent analyses of four real data sets, however, the performance order of the two strategies was clearly reversed. To clarify this situation, Schepers et al. (2006) ran a number of model checks on their real data. Those checks revealed clear violations of the implicit stochastic assumptions underlying loss function (15) (which otherwise had also been relied on in the generation of synthetic data in the first simulation study).

More in particular, the model checks revealed clear violations of the independence assumption for the error terms,

$$e_{i_1 \ldots i_n \ldots i_N} \overset{iid}{\sim} N(0, \sigma^2).$$

Schepers et al. (2006) subsequently ran a second simulation study, with the use of nonzero error covariances in the generation of the synthetic data. In this second simulation study, the performance reversal at the advantage of the second algorithmic strategy, as observed in the analysis of the real data, was replicated. As a tentative explanation of these findings, we conjecture that Strategy 1 may work better for problems in which not only the loss function at hand, but also the likelihood of the "true" underlying stochastic model is separable (possibly upon ignoring the rank deficiency issue); if separability of the likelihood, however, gets lost due to the presence of covariance terms, Strategy 2 may work better. Anyhow, this example should make clear that one should be careful in drawing fast conclusions about the relative performances of the two generic algorithmic strategies as outlined above.

Acknowledgments

Work on this paper has been supported by the Fund for Scientific Research—Flanders (project G.0146.06) and by the Research Fund of K.U. Leuven (GOA/2005/04 and EF/05/007). The authors gratefully acknowledge Henk Kiers and Eva Ceulemans for their useful comments on a previous version of this manuscript.

References

Bock, H.-H., 1979. Simultaneous clustering of objects and variables. In: Tomassone, R. (Ed.), Analyse des données et informatique. INRIA, Le Chesnay, pp. 187–204.

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

Bro, R., 1998. Multi-way analysis in the food industry. Models, algorithms and applications. Unpublished doctoral dissertation. University of Amsterdam, Amsterdam.

Carroll, J.D., Arabie, P., 1980. Multidimensional scaling. Annu. Rev. Psychol. 31, 607–649.

Carroll, J.D., Chaturvedi, A., 1995. A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data. In: Luce, R.D., D'Zmura, M., Hoffman, D.D., Iverson, G.J., Romney, A.K. (Eds.), Geometric Representations of Perceptual Phenomena. Erlbaum, Mahwah, pp. 295–318.

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

Ceulemans, E., Van Mechelen, I., Leenen, I., 2003. Tucker3 hierarchical classes analysis. Psychometrika 68, 413–433.

Ceulemans, E., Van Mechelen, I., Kuppens, P., 2004. Adapting the formal to the substantive: constrained Tucker3-HICLAS. J. Classification 21, 19–50.

Coombs, C.H., 1964. A Theory of Data. University of Michigan Press, Ann Arbor.

De Boeck, P., Rosenberg, S., 1988. Hierarchical classes: model and data analysis. Psychometrika 53, 361–381.

DeSarbo, W.S., Carroll, J.D., 1985. Three-way metric unfolding via alternating weighted lesat squares. Psychometrika 50, 275–300.

De Soete, G., Heiser, W.J., 1993. A latent class unfolding model for analyzing single stimulus preference ratings. Psychometrika 58, 545–565.

Eckes, T., Orlik, P., 1994. Three-mode hierarchical cluster analysis of three-way three-mode data. In: Bock, H.-H., Lenski, W., Richter, M.M. (Eds.), Information Systems and Data Analysis: Prospects Foundations Applications. Springer, Heidelberg, pp. 217–225.

Furnas, G.W., 1980. Objects and their features: the metric analysis of two-class data. Unpublished doctoral dissertation. Stanford University, Stanford.

Gaul, W., Schader, M., 1996. A new algorithm for two-mode clustering. In: Bock, H.-H., Polasek, W. (Eds.), Data Analysis and Information Systems. Springer, Heidelberg, pp. 15–23.

Guttman, L., 1944. A basis for scaling qualitative data. Amer. Sociological Rev. 9, 139–150.

Heiser, W.J., Groenen, P.J.F., 1997. Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. Psychometrika 62, 63–83.

Kiers, H.A.L., 2000. Towards a standardized notation and terminology in multiway analysis. J. Chemometrics 14, 105–122.

Kiers, H.A.L., 2004. Clustering all three modes of three-mode data: computational possibilities and problems. In: Antoch, J. (Ed.), COMPSTAT Proceedings in Computational Statistics. Springer, Heidelberg, pp. 303–313.

Kiers, H.A.L., Smilde, A.K., 1998. Constrained three-mode factor analysis as a tool for parameter estimation with second-order instrumental data. J. Chemometrics 12, 125–147.

Kroonenberg, P.M., 1983. Three-mode Principal Component Analysis: Theory and Applications. DSWO, Leiden.

Leenen, I., Van Mechelen, I., De Boeck, P., Rosenberg, S., 1999. INDCLAS: individual differences hierarchical classes analysis. Psychometrika 64, 9–24.

Leenen, I., Van Mechelen, I., De Boeck, P., 2001. Models for ordinal hierarchical classes analysis. Psychometrika 66, 389–404.
McLachlan, G.J., Chang, S.U., 2004. Mixture modelling for cluster analysis. Stat. Meth. Med. Res. 13, 347–361.
Rocci, R., Vichi, M., 2005. Three-mode clustering of a three-way data set. In: Vichi, M., Monari, P., Mignani, S., Montanari, A. (Eds.), New Developments in Classification and Data Analysis. Springer, Berlin.
Schepers, J., Van Mechelen, I., Ceulemans, E., 2006. Three-mode partitioning. Comput. Stat. Data Anal. 51, 1623–1642.
Tucker, L.R., 1964. The extension of factor analysis to three-dimensional matrices. In: Frederiksen, N., Gulliksen, H. (Eds.), Contributions to Mathematical Psychology. Erlbaum, Hillsdale, pp. 109–127.
Van Deun, K., Marchal, K., Heiser, W.J., Engelen, K., Van Mechelen, I., 2007. Joint mapping of genes and conditions via multidimensional unfolding analysis. BMC Bioinformatics, submitted for publication.
Van Mechelen, I., De Boeck, P., Rosenberg, S., 1995. The conjunctive model of hierarchical classes. Psychometrika 60, 505–521.
Van Mechelen, I., Bock, H.-H., De Boeck, 2004. Two-mode clustering methods: a structural overview. Stat. Meth. Med. Res. 13, 363–394.
Van Mechelen, I., Lombardi, L., Ceulemans, E., in press. Hierarchical classes modeling of rating data. Psychometrika.
Vichi, M., Kiers, H.A.L., 2001. Factorial *k*-means analysis for two-way data. Comput. Stat. Data Anal. 37, 49–64.