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.


*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.