Computational Statistics and Data Analysis 53 (2009) 1086–1098

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Simultaneous analysis of coupled data blocks differing in size: A comparison of two weighting schemes

Tom Wilderjans *, Eva Ceulemans, Iven Van Mechelen

Katholieke Universiteit Leuven, Department of Psychology, Tiensestraat 102, Leuven, Belgium

ARTICLE INFO

Article history:
Received 8 January 2008
Received in revised form 26 September 2008
Accepted 29 September 2008
Available online 9 October 2008

ABSTRACT

Research questions in several research domains imply the simultaneous analysis of different blocks of information that pertain to the same research objects. In personality psychology, for example, to study the relation between individual differences in behavior and cognitive-affective units that can account for these differences, two types of information pertaining to the same set of persons need to be analyzed simultaneously: (1) information about the situation-specific behavior profile of these persons, and (2) information about the cognitive-affective units these persons exhibit. When dealing with such coupled data blocks (i.e., different N-way N-mode data blocks that have one or more modes in common) it often happens that one data block is much larger in size than the other(s). In this case, the question arises whether the data entries or the data blocks should be considered as the units of information, in order to disclose the true structure underlying the coupled data blocks. To answer this question, two weighting schemes are compared that are obtained by applying weights in the overall objective function that is to be optimized in the data analysis, with each weight indicating the extent to which the corresponding data block influences the integrated analysis. In a simulation study it is showed that weighting the different data blocks such that each data entry influences the analysis to the same extent (i.e., data entries as units of information) outperforms a weighting scheme in which each data block has an equal influence on the analysis (i.e., data blocks as units of information). This superior performance is demonstrated for two global models for coupled data consisting of a three-way three-mode data block and a two-way two-mode data block that have one mode in common: (1) a multiway multiblock component model for coupled real-valued data, and (2) a simultaneous clustering model for coupled binary data.

1. Introduction

As a consequence of our information society, researchers in many areas of science face an increasing flow of information that pertains both to the amount and the complexity of the data. In particular, this often implies that different pieces of information are collected with respect to the same research objects, with distinct pieces possibly being of a different nature and emanating from different sources. Such data, which in the remainder will be referred to as coupled data, can be defined as a set of N-way N-mode data blocks, with each data block having at least one mode in common with at least one other data block. Data of this type occur frequently in many areas of science. As an example, one may think of research in contextualized personality psychology in which one may wish to describe individual differences in situation-specific

* Corresponding address: Katholieke Universiteit Leuven, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium. Tel.: +32 16 326108; fax: +32 16 325993.

E-mail address: tom.wilderjans@psy.kuleuven.be (T. Wilderjans).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.09.031

behavior and to simultaneously uncover the psychological processes that account for these individual differences. For this purpose, for a set of persons the following information may be gathered: (1) information on the degree to which each of the persons displays a set of behaviors in a set of situations, and (2) information on the extent to which each person makes use of a set of cognitive-affective units. This results in a data block couple consisting of a three-way three-mode person by situation by behavior data block and a two-way two-mode person by cognitive-affective unit data block that have one mode (i.e., the person mode) in common.

When analyzing coupled data, a researcher is faced with the challenging task to make an appropriate synthesis of the different pieces of information that pertain to the same research objects. Since each piece may contain information about different aspects of the objects under study, the true structure underlying these objects, with this true structure being the same for all pieces of information, can best be disclosed by integrating all the available information. From a data-analytic viewpoint, this implies that a global model is needed in which the different data blocks, one block per piece of information, are analyzed simultaneously. In the present paper, in this regard, global models will be considered that consist of different submodels, one for each data block, with each submodel implying a (dimensional or categorical) quantification of all modes of the corresponding data block (Van Mechelen and Schepers, 2007). Further, only global models, consisting of different submodels, will be considered in which each common mode of the coupled data is represented by a single quantification, which is the same for all submodels of the global model that mode belongs to. Note that such a global model may also be able to reveal the structure underlying the common mode(s) when, with respect to that common mode(s), the information in one piece can be captured by fewer dimensions or categories than the information in another piece. In that case, some dimensions or categories of the (single) quantification for the common mode will account in part for the information in one piece, while not accounting for any information in another piece. Although, at the level of the model, such a situation may be dealt with, it still has to be demonstrated (e.g., in a simulation study) how well the structure underlying the common mode is disclosed in such a situation. To fit such a global model to a given coupled data set at hand, an integrated strategy will be followed in which the global model is fitted to the coupled data set at once. As a result, to derive the quantification of a common mode, the information in all data blocks in which the mode in question shows up, is used. As such, information on the common modes as stemming from different sources is combined (*data fusion*). Applied to the contextualized personality psychology example, a global model consisting of a three-way three-mode and a two-way two-mode submodel may be fitted simultaneously to the person by situation by behavior data block and the person by cognitive-affective unit data block, respectively, with a common quantification of the person mode being estimated on the basis of the information in both data blocks. Note that this strategy implies the optimization of a global loss function, which, for example, can be a weighted sum of partial loss functions, one for each submodel.

When using a global loss function that is a weighted sum of partial loss functions, two approaches may be used to determine the weights. First, the weights may be determined a priori based on considerations like the amount of influence each piece of information should have on the analysis. Note that standardizing the data blocks is equivalent to an appropriate (a priori) weighting of these blocks. Second, the weights could be estimated during the data-analytic process. In this regard, Smilde and Kiers (1999) propose, in the context of a multiway covariates regression model, to estimate the weights based on cross-validation, with the sum of the weights being restricted to equal one.

In this paper, the weights will be determined a priori in order to deal with coupled data sets for which one data block is much larger than the other(s). Such data occur often in psychology. For instance, in the contextualized personality psychology example, the person by situation by behavior data block is often much larger than the person by cognitive-affective unit data block. As a consequence, when considering each data entry as a unit of information, the largest block contains more information with respect to (the underlying structure of) the common mode than the smaller block(s). It is also possible, however, to take each data block as a unit of information, implying that all data blocks contribute equally. One question that arises is, irrespective of the (type of) structure underlying the coupled data, whether the data entries or the data blocks should be considered as the units of information, in order to (best) disclose this true underlying structure. When the data entries should be considered as the units of information, it is recommended to let each data entry influence the derivation of the underlying structure of the common mode to the same extent, while when the data blocks should be considered as the units of information, it is preferred to give the same amount of influence to each data block. Note that in the latter case the data entries of the larger data block influence the derivation of the common mode to a smaller extent than data entries from the smaller data block(s). To manipulate the amount of influence of the data elements/blocks on the derivation of the common mode, in the analysis weights could be applied to the partial loss functions pertaining to the different data blocks, with each weight indicating the extent to which the corresponding data block influences the integrated analysis (i.e., the estimation of the parameters of the common mode). Both cases then boil down to two different weighting schemes: weighting the different data blocks such that (1) each data entry influences the analysis to the same extent (first case), or (2) each data block has an equal influence on the analysis (second case). In the present paper, it will be studied, by means of a simulation study, which weighting scheme is preferable in terms of uncovering the true structure underlying (the common modes of) a coupled data set. This weighting issue will be dealt with within the context of two global models for coupled data consisting of a three-way three-mode data block and a two-way two-mode data block that have one mode in common: (1) a multiway multiblock component model for the case of coupled real-valued data (Smilde et al., 2000), and (2) a simultaneous clustering model, called CHIC, for the case of coupled binary data (Wilderjans et al., in press).

The remainder of this paper is organized in three main sections: In Section 2, the weighting issue is treated for a multiway multiblock component model for coupled real-valued data. In Section 3, this issue is handled for the CHIC model for coupled

binary data. In both sections, first the corresponding global model is presented, along with the loss function to be optimized. Next, a simulation study to investigate which weighting scheme performs best is reported. In Section 4, the results of both simulation studies are compared and discussed.

2. First study: The multiway multiblock component model

2.1. Multiway multiblock component analysis

2.1.1. Model

The specific multiway multiblock component model under study (for more information about the family of multiway multiblock component models, see Smilde et al. (2000)), approximates an $I \times J \times K$ object by attribute by source real-valued data block $\underline{\mathbf{D}}^1$ and an $I \times L$ object by covariate real-valued data block $\mathbf{D}^2$ by an $I \times J \times K$ real-valued model block $\underline{\mathbf{M}}^1$ and an $I \times L$ real-valued model block $\mathbf{M}^2$, respectively. In this model, (1) $\underline{\mathbf{M}}^1$ can be decomposed according to a rank P CANDECOMP/PARAFAC (sub)model (Harshman, 1970; Carroll and Chang, 1970), while (2) $\mathbf{M}^2$ can be decomposed into a rank P PCA (sub)model; hence, the decomposition of the model matrices $\underline{\mathbf{M}}^1$ and $\mathbf{M}^2$ reads as follows:

$$\mathbf{M}_a^1 = \mathbf{A}\mathbf{G}(\mathbf{C} \otimes \mathbf{B}_1)^{\mathrm{T}}, \tag{1}$$

$$\mathbf{M}^2 = \mathbf{A}\mathbf{B}_2^{\mathrm{T}}, \tag{2}$$

where the $I \times JK$ matrix $\mathbf{M}_a^1$ and the $P \times PP$ matrix $\mathbf{G}$ are the matricized versions (Kiers, 2000) of $\underline{\mathbf{M}}^1$ and the three-way $P \times P \times P$ superidentity matrix $\underline{\mathbf{G}}$ (i.e., $g_{rst} = 1$ iff $r = s = t = p$ and $g_{rst} = 0$ otherwise), respectively, $\otimes$ denotes the Kronecker product, and T denotes matrix transposition. The matrices $\mathbf{A}$ ($I \times P$), $\mathbf{B}_1$ ($J \times P$), $\mathbf{C}$ ($K \times P$), and $\mathbf{B}_2$ ($L \times P$) are the real-valued component matrices for the objects, attributes, sources, and covariates, respectively. Note that the submodels are linked to each other by constraining the common object component matrix $\mathbf{A}$ to be the same in both submodels.

As an aside one may note that the family of multiway multiblock component models is closely related to the family of multiway covariates regression models (Smilde and Kiers, 1999). More specifically, both model families differ in one respect only, that is, in the role that is assigned to the different data blocks. In a multiway covariates regression model, the different data blocks have two distinctive roles (i.e., one block serves as a predictor block while the other block serves as a criterion block), while in a multiway multiblock component model, to the contrary, the different data blocks are exchangeable in terms of conceptual status. Note further that in Smilde et al. (2000) the family of multiway covariates regression models is called the family of multiway multiblock regression models.

2.1.2. Loss function

The aim of a multiway multiblock component analysis in rank P of an $I \times J \times K$ real-valued data block $\underline{\mathbf{D}}^1$ and an $I \times L$ real-valued data block $\mathbf{D}^2$ is to estimate an $I \times J \times K$ real-valued model block $\underline{\mathbf{M}}^1$ and an $I \times L$ real-valued model block $\mathbf{M}^2$ such that (1) the value of the loss function

$$f(\underline{\mathbf{M}}^1, \mathbf{M}^2) = \alpha \|\underline{\mathbf{D}}^1 - \underline{\mathbf{M}}^1\|^2 + \beta \|\mathbf{D}^2 - \mathbf{M}^2\|^2 \tag{3}$$

is minimized, with $\|\cdots\|$ denoting the Euclidean (matrix) norm, and (2) $\underline{\mathbf{M}}^1$ and $\mathbf{M}^2$ can be represented by a rank P CANDECOMP/PARAFAC model and a rank P PCA model, respectively, with an identical object component matrix $\mathbf{A}$. The weights α and β indicate the importance of a correct reconstruction of $\underline{\mathbf{D}}^1$ and $\mathbf{D}^2$, respectively. Note, however, that the extent to which each data block influences the loss function in (3) is also determined by the scale on which each data block is measured. To remove possible differences in scale between both data blocks, in this paper it is assumed that each data block is normalized by dividing each entry of that data block by the range of the corresponding data block (i.e., $\underline{\mathbf{D}}^1 = \frac{\underline{\mathbf{D}}^1}{\max(\underline{\mathbf{D}}^1) - \min(\underline{\mathbf{D}}^1)}$ and $\mathbf{D}^2 = \frac{\mathbf{D}^2}{\max(\mathbf{D}^2) - \min(\mathbf{D}^2)}$). When, however, the variables within each block are measured on a different scale, the normalization should be applied variablewise instead of blockwise. Note further that the loss function in (3) is a generic loss function with the loss function presented in Smilde et al. (2000) being a specific instance thereof (i.e., $\alpha = \frac{\alpha^*}{\|\underline{\mathbf{D}}^1\|^2}$ and $\beta = \frac{1-\alpha^*}{\|\mathbf{D}^2\|^2}$ with α^* taking a value between 0 and 1).

To estimate the parameters of the multiway multiblock component model, an alternating least squares (ALS) algorithm is used (Kroonenberg and De Leeuw, 1980) in which the component matrices $\mathbf{A}$, $\mathbf{B}_1$, $\mathbf{C}$, and $\mathbf{B}_2$ are, alternatingly re-estimated conditionally upon the other component matrices until no improvement in the loss function is observed (see Smilde and Kiers (1999)).

2.2. Simulation study

2.2.1. Problem

In this section, a simulation study is set up to compare the performance, with respect to the disclosure of the true structure underlying a coupled data set, of the two weighting schemes that were proposed in the introduction: i.e., one scheme implying an equal influence of each data block on the global loss function and the other one implying an equal influence of each data entry on the global loss function. To ensure equal influence of the data blocks, the partial loss functions of the global loss function are weighted by the inverse of the size of the corresponding data block (i.e., $\alpha = \frac{1}{IJK}$ and $\beta = \frac{1}{IL}$); in this way, the following block-based global loss function is obtained:

$$\frac{\|\underline{\mathbf{D}}^1 - \underline{\mathbf{M}}^1\|^2}{IJK} + \frac{\|\mathbf{D}^2 - \mathbf{M}^2\|^2}{IL}. \tag{4}$$

To obtain equal influence of each data entry, each partial loss function is weighted by 1 (i.e., $\alpha = 1$ and $\beta = 1$); as such, the following entry-based global loss function is obtained:

$$\|\underline{\mathbf{D}}^1 - \underline{\mathbf{M}}^1\|^2 + \|\mathbf{D}^2 - \mathbf{M}^2\|^2. \tag{5}$$

To address the question which of these two weighting schemes best uncovers the true structure underlying a coupled data set, the recovery performance of both weighting schemes will be evaluated both at the level of the model matrices and at the level of the component matrices. To this end, three types of real-valued $I \times J \times K$ and $I \times L$ block couples must be distinguished: (1) the (known) true block couple ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$), of rank P, that can be decomposed into $\mathbf{A}^{(T)}$, $\mathbf{B}_1^{(T)}$, $\mathbf{C}^{(T)}$, and $\mathbf{B}_2^{(T)}$, (2) the data block couple ($\underline{\mathbf{D}}^1$, $\mathbf{D}^2$) that is obtained by adding noise to ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$), and (3) the model block couple ($\underline{\mathbf{M}}^1$, $\mathbf{M}^2$), also of rank P, that is obtained by applying the multiway multiblock component algorithm in rank P to the data block couple ($\underline{\mathbf{D}}^1$, $\mathbf{D}^2$) and that can be decomposed into $\mathbf{A}^{(M)}$, $\mathbf{B}_1^{(M)}$, $\mathbf{C}^{(M)}$, and $\mathbf{B}_2^{(M)}$. For each weighting scheme it then can be determined to which extent the entries of the true block couple are recovered by comparing the entries of the model block couple (in $\underline{\mathbf{M}}^1$ and $\mathbf{M}^2$) to the entries of the true block couple (in $\underline{\mathbf{T}}^1$ and $\mathbf{T}^2$). Further, the recovery of the true component matrices can be studied by comparing the component matrices as obtained from the algorithm (i.e., $\mathbf{A}^{(M)}$, $\mathbf{B}_1^{(M)}$, $\mathbf{C}^{(M)}$, and $\mathbf{B}_2^{(M)}$) to the true component matrices (i.e., $\mathbf{A}^{(T)}$, $\mathbf{B}_1^{(T)}$, $\mathbf{C}^{(T)}$, and $\mathbf{B}_2^{(T)}$).

In the remainder, first, the design and the procedure of the simulation study will be outlined. Next, the simulation results with respect to the recovery performance at the level of the model matrices (i.e., the entries of the true block couple) and the level of the component matrices, will be presented.

2.2.2. Design and procedure

Three factors were systematically manipulated in a completely randomized three-factorial design, all factors considered random:

(a) the *Array/total ratio*, r, of the size of $\underline{\mathbf{T}}^1$ ($\underline{\mathbf{D}}^1$, $\underline{\mathbf{M}}^1$) to the total size $\underline{\mathbf{T}}^1 + \mathbf{T}^2$ ($\underline{\mathbf{D}}^1 + \mathbf{D}^2$, $\underline{\mathbf{M}}^1 + \mathbf{M}^2$, respectively):

$$r = \frac{I \times J \times K}{(I \times J \times K) + (I \times L)}. \tag{6}$$

This factor was manipulated at four levels: .50, .90, .95, .99. Note that the absolute size of $\underline{\mathbf{T}}^1$ ($\underline{\mathbf{D}}^1$, $\underline{\mathbf{M}}^1$) was kept constant at the level of 27 000 entries;

(b) the *Dominance in size*, d, of the common mode over the other modes of $\underline{\mathbf{T}}^1$ ($\underline{\mathbf{D}}^1$, $\underline{\mathbf{M}}^1$). This factor was manipulated at three levels: one level with the common mode being the largest mode ($50 \times 20 \times 27$), one level with the common mode being as large as the other modes ($30 \times 30 \times 30$), and one level with the common mode being the smallest mode ($20 \times 50 \times 27$);

(c) the *Error level*, ε, defined as the (expected) ratio of the sum of squares of the error (i.e., the difference between the data and the truth) to the sum of squares of the data:

$$\varepsilon = E\left[\frac{\|\underline{\mathbf{D}}^1 - \underline{\mathbf{T}}^1\|^2}{\|\underline{\mathbf{D}}^1\|^2}\right] = E\left[\frac{\|\mathbf{D}^2 - \mathbf{T}^2\|^2}{\|\mathbf{D}^2\|^2}\right], \tag{7}$$

with $E[.]$ denoting the expected value. This factor was manipulated at four levels: .00, .25, .50, .75.

For each combination of an Array/total ratio, a Dominance, and an Error level, rank 4 true component matrices $\mathbf{A}^{(T)}$, $\mathbf{B}_1^{(T)}$, $\mathbf{C}^{(T)}$, and $\mathbf{B}_2^{(T)}$ were constructed by independently drawing entries from a standard normal distribution. Note that the True rank of the multiway multiblock component model for ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$) was not manipulated here, because in a pilot study it appeared that this factor is not influencing the results. Next, the true block couple ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$) was computed by combining the true component matrices $\mathbf{A}^{(T)}$, $\mathbf{B}_1^{(T)}$, $\mathbf{C}^{(T)}$, and $\mathbf{B}_2^{(T)}$ by the CANDECOMP/PARAFAC and PCA decomposition rules (1) and (2). Further, for each true block couple ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$), a data block couple ($\underline{\mathbf{D}}^1$, $\mathbf{D}^2$) was constructed by the following procedure. First, an $I \times J \times K$ error matrix $\underline{\mathbf{E}}^1$ and an $I \times L$ error matrix $\mathbf{E}^2$ were created by drawing independently entries from a

Table 1
Mean recovery of true block couple entries for analyses with block-based and entry-based weighting schemes, calculated for the global model as a whole and for the separate submodels.

	Weighting scheme	
	Block-based	Entry-based
Global model	0.00039	0.00016
Submodels		
PCA	0.00242	0.00106
CANDECOMP/PARAFAC	0.00025	0.00004

univariate standard normal distribution. Second, both error matrices were scaled (i.e., by multiplying $\underline{\mathbf{E}}^1$ by $\sqrt{\frac{\|\underline{\mathbf{T}}^1\|^2}{\|\underline{\mathbf{E}}^1\|^2}}$ and $\mathbf{E}^2$ by $\sqrt{\frac{\|\mathbf{T}^2\|^2}{\|\mathbf{E}^2\|^2}}$) to ensure that $\|\underline{\mathbf{E}}^1\|^2 = \|\underline{\mathbf{T}}^1\|^2$ and $\|\mathbf{E}^2\|^2 = \|\mathbf{T}^2\|^2$. Next, $\underline{\mathbf{E}}^1$ and $\mathbf{E}^2$ were multiplied by a scalar $\sqrt{c}$, resulting in $\|\underline{\mathbf{E}}^1\|^2 = c\|\underline{\mathbf{T}}^1\|^2$ and $\|\mathbf{E}^2\|^2 = c\|\mathbf{T}^2\|^2$, and the obtained $\underline{\mathbf{E}}^1$ and $\mathbf{E}^2$ were added to $\underline{\mathbf{T}}^1$ and $\mathbf{T}^2$, respectively. The value for c was set equal to 0, .333, 1, and 3 for the conditions with $\varepsilon = 0$, .25, .50, and .75, respectively, to ensure that the (expected) ratio of the sum of squares of the error to the sum of squares of the data being equal to 0, .25, .50, and .75. Note that $\varepsilon = \frac{c}{c+1}$. This data generation procedure was repeated 20 times to obtain 20 replications per cell. As a consequence, 20 (replications) $\times$ 4 (Array/total ratio) $\times$ 3 (Dominance) $\times$ 4 (Error level) $=$ 960 different true block couples ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$) and data block couples ($\underline{\mathbf{D}}^1$, $\mathbf{D}^2$) were obtained. Finally, two rank 4 multiway multiblock component analyses, one with the block-based loss function and one with the entry-based loss function, were applied to each data block couple ($\underline{\mathbf{D}}^1$, $\mathbf{D}^2$). In order to lower the risk of ending in a local optimum, for each analysis 50 runs of the ALS algorithm were performed (one run with a rational start, the other runs with a random start) and the run resulting in the lowest value on the loss function was retained. To obtain a rational start for $\mathbf{A}$ ($\mathbf{B}_1$), PCA was performed on $[\mathbf{D}_a^1|\mathbf{D}^2]$ ($\mathbf{D}_b^1$), with $\mathbf{D}_a^1$ ($I \times JK$) and $\mathbf{D}_b^1$ ($J \times KI$) being the matricized versions of $\underline{\mathbf{D}}^1$ (Kiers, 2000), and | denoting matrix concatenation.

2.2.3. Results

The recovery performance of both weighting schemes is studied both at the level of the model matrices and at the level of the (individual) component matrices. At the level of the model matrices, the extent to which the entries of the true block couple are recovered, was determined by computing the mean sum of squared differences between the model block couple ($\underline{\mathbf{M}}^1$, $\mathbf{M}^2$) and the true block couple ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$). Further, at the level of the individual component matrices, recovery performance was measured by computing for each component matrix a *goodness-of-recovery* statistic (*GOR*). In particular, for the common component matrix $\mathbf{A}$, *GOR* was defined as follows:

$$GOR_{\mathbf{A}} = \frac{\sum_{p=1}^{P} |\phi(\mathbf{a}_p^{(T)}, \mathbf{a}_p^{(M)})|}{P}, \tag{8}$$

with ϕ being the Tucker phi-coefficient (Tucker, 1951), $\mathbf{a}_p^{(T)}$ ($\mathbf{a}_p^{(M)}$) the pth component of the true (obtained) common component matrix $\mathbf{A}^{(T)}$ ($\mathbf{A}^{(M)}$), and P the rank of the model (i.e., the number of components), which always equals 4 here. This *GOR* statistic takes values between 0 (no recovery at all) and 1 (perfect recovery). A similar *GOR* statistic was computed for $\mathbf{B}_1$, $\mathbf{C}$, and $\mathbf{B}_2$. Note that by taking the absolute value of the Tuckers phi-coefficient, the reflectional freedom of the components of a multiway multiblock component solution is dealt with. Note further that to account for the (joint) permutational freedom of the components in $\mathbf{A}^{(M)}$, $\mathbf{B}_1^{(M)}$, $\mathbf{C}^{(M)}$, and $\mathbf{B}_2^{(M)}$, the permutation was chosen that maximizes the following criterion:

$$\sum_{p=1}^{P} [|\phi(\mathbf{a}_p^{(T)}, \mathbf{a}_p^{(M)})| \times |\phi(\mathbf{b}_{1_p}^{(T)}, \mathbf{b}_{1_p}^{(M)})| \times |\phi(\mathbf{c}_p^{(T)}, \mathbf{c}_p^{(M)})| \times |\phi(\mathbf{b}_{2_p}^{(T)}, \mathbf{b}_{2_p}^{(M)})|]. \tag{9}$$

To compute an overall *GOR* statistic, a combined measure (*cGOR*) was constructed by calculating the weighted average of the *GOR* of the individual component matrices, weighted by the number of elements of the corresponding mode.

In Table 1, for both weighting schemes, the mean sum of squared differences between the model block couple ($\underline{\mathbf{M}}^1$, $\mathbf{M}^2$) and the true block couple ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$) is displayed, both for the global model, consisting of the two submodels, as a whole (i.e., this was obtained by taking all elements of both blocks together in one superblock, implying each element being equally important) and for the two submodels separately. For the global model as a whole, the entry-based weighting scheme outperforms the block-based weighting scheme, with the former yielding a smaller sum of squared differences than the latter one. Also for the submodels this appears to be the case, with the difference in recovery performance being most pronounced for the two-way submodel (i.e., the PCA model).

In Table 2, for both weighting schemes, the mean *GOR* value is displayed, in terms of the combined measure (*cGOR*) and the *GOR* measure for the separate component matrices. On average, *cGOR* appears to be larger for the entry-based loss function than for the block-based loss function, implying that the weighting scheme in which each data entry has an equal

Table 2
Mean recovery of true component matrices (*GOR*) for analyses with block-based and entry-based weighting schemes, calculated as a whole and for the separate component matrices.

	Weighting scheme	
	Block-based	Entry-based
Whole model (*cGOR*)	0.9440	0.9643
Separate component matrices		
A	0.9515	0.9968
$\mathbf{B}_1$	0.9955	0.9968
C	0.9960	0.9974
$\mathbf{B}_2$	0.8976	0.9287

influence on the global loss function outperforms the weighting scheme in which each data block influences the analysis to the same extent. Further, a similar result was obtained at the level of the individual component matrices, with the difference in recovery performance being the largest for the common component matrix **A**.

To further compare the recovery performance of both weighting schemes, the following statistic was defined:

$$cGOR^{diff} = cGOR^{entry} - cGOR^{block}. \tag{10}$$

Note that an analogous *diff* statistic may also be defined for the sum of squared differences between $(\underline{\mathbf{M}}^1, \mathbf{M}^2)$ and $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$. Because the pattern of the results for this statistic is the same as for the $cGOR^{diff}$ statistic, further, only results for the latter statistic will be reported. To gain insight in how the difference in recovery performance between both weighting schemes varies with the manipulated characteristics of the data, in Fig. 1, the mean $cGOR^{diff}$ value is displayed as a function of the Array/total ratio, the Error level, and the Dominance in size. Further, an analysis of variance with $cGOR^{diff}$ as dependent variable was performed in which it turned out that all main and interaction effects of these data characteristics are significant at .05 level. Only discussing effects with an intraclass correlation (Haggard, 1958; Kirk, 1982) larger than or equal to .10 ($\hat{\rho}_I \geq .10$), this analysis revealed that the entry-based loss function especially outperforms the block-based loss function when the Error level increases ($\hat{\rho}_I = .13$) and the Array/total ratio increases ($\hat{\rho}_I = .23$). Both main effects, however, are qualified by an Error level by Array/total ratio interaction ($\hat{\rho}_I = .35$) and an Array/total ratio by Dominance interaction ($\hat{\rho}_I = .11$): The two effects are stronger when the three-way data block is much larger than the two-way data block and when the common mode is the most sizeable one, respectively. These interactions, however, in turn, are qualified by an Error level by Array/total ratio by Dominance interaction ($\hat{\rho}_I = .10$): the pattern of the Error level by Array/total ratio interaction is most pronounced when the common mode is the larger mode of the three-way data block. Note that when $r = .50$, both weighting schemes perform equally well, because in that condition both weighting schemes coincide (i.e., $IJK = IL$ and thus the entry-based weighting scheme is equivalent to the block-based one in that both blocks are weighted equally).

3. Second study: The CHIC model

3.1. CHIC analysis

3.1.1. Model

The CHIC model (Wilderjans et al., in press) approximates an $I \times J \times K$ object by attribute by source binary data block $\underline{\mathbf{D}}^1$ and an $I \times L$ object by covariate binary data block $\mathbf{D}^2$ by an $I \times J \times K$ binary model block $\underline{\mathbf{M}}^1$ and an $I \times L$ binary model block $\mathbf{M}^2$, respectively. With respect to this model, (1) $\underline{\mathbf{M}}^1$ and $\mathbf{M}^2$ are modeled with an INDCLAS (sub)model (Leenen et al., 1999) and a HICLAS (sub)model (De Boeck and Rosenberg, 1988) of rank P, respectively, and (2) both submodels are linked to one another by restricting the common object component matrix **A** to be the same in both submodels. The model matrices $\underline{\mathbf{M}}^1$ and $\mathbf{M}^2$ can therefore be reconstructed by the following decomposition rules respectively:

$$m^1_{ijk} = \bigoplus_{p=1}^{P} a_{ip} b^1_{jp} c_{kp}, \tag{11}$$

$$m^2_{il} = \bigoplus_{p=1}^{P} a_{ip} b^2_{lp}, \tag{12}$$

where $\bigoplus$ denotes a Boolean sum, P indicates the rank of the model, and a_{ip}, b^1_{jp}, c_{kp}, and b^2_{lp} denote the entries of **A** ($I \times P$), $\mathbf{B}^1$ ($J \times P$), **C** ($K \times P$), and $\mathbf{B}^2$ ($L \times P$), the binary component matrices for the objects, attributes, sources, and covariates, respectively. The P columns of **A**, $\mathbf{B}^1$, **C**, and $\mathbf{B}^2$, which are called components (or bundles), define P, possibly overlapping, clusters of objects, attributes, sources, and covariates, respectively. Note that the decomposition rules in (11) and (12) imply a one-to-one correspondence among the respective object, attribute, source, and covariate components.

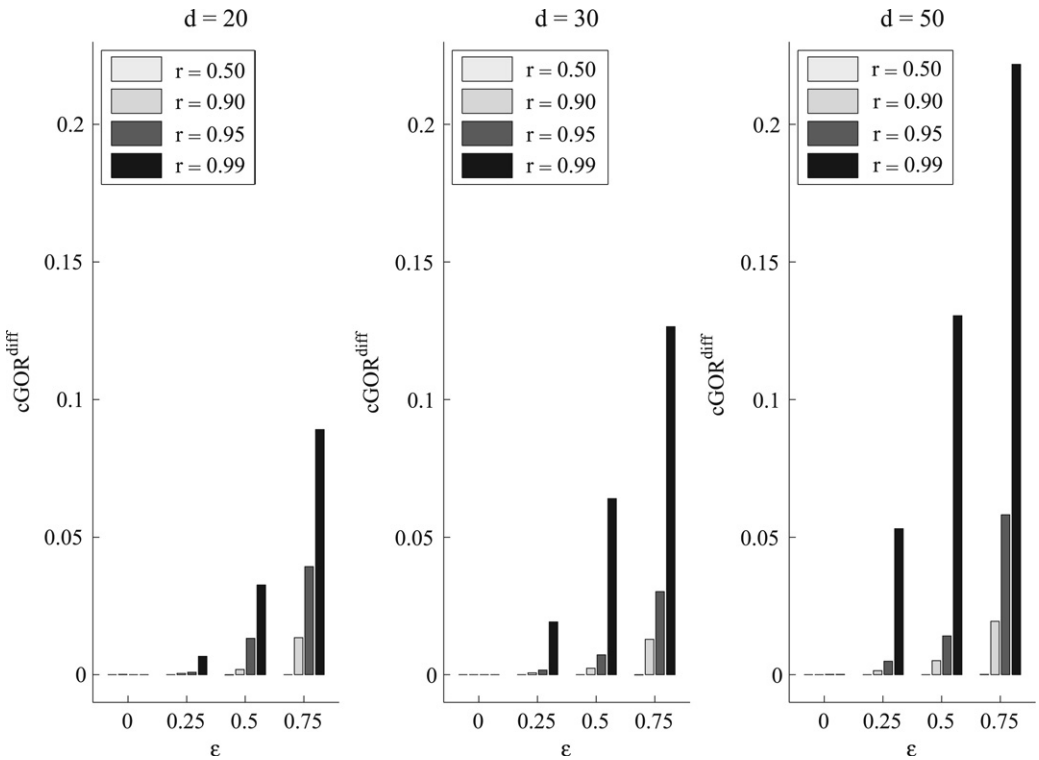

Fig. 1. Mean $cGOR^{diff}$ value as a function of the Array/total ratio (r), the Error level (ε), and the Dominance in size (d).

3.1.2. Loss function

Given a binary data block couple ($\underline{\mathbf{D}}^1$, $\mathbf{D}^2$) and a rank P, the aim of a CHIC analysis is to estimate a binary model block couple ($\underline{\mathbf{M}}^1$, $\mathbf{M}^2$) such that (1) the value of the loss function

$$f(\underline{\mathbf{M}}^1, \mathbf{M}^2) = \alpha\|\underline{\mathbf{D}}^1 - \underline{\mathbf{M}}^1\|^2 + \beta\|\mathbf{D}^2 - \mathbf{M}^2\|^2 \tag{13}$$

is minimized, and (2) $\underline{\mathbf{M}}^1$ and $\mathbf{M}^2$ can be represented by a rank P INDCLAS model and a rank P HICLAS model, respectively, with a common object component matrix $\mathbf{A}$. The coefficients α and β again denote the weight of the corresponding data block in the analysis. Note that, whereas with real-valued data possible scale differences between data blocks had to be taken into account, scale differences are not an issue here because of the binary nature of the data. Note further that the CHIC loss function presented in Wilderjans et al. (in press) is a special case of the generic loss function in (13), with $\alpha = 1$ and $\beta = 1$.

The parameters of the CHIC algorithm may be estimated by means of a simulated annealing algorithm (a full description of this algorithm can be found in Wilderjans et al. (in press)).

3.2. Simulation study

3.2.1. Problem

A simulation study is performed to determine, within the context of the CHIC model, which weighting scheme best discloses the true structure underlying a coupled data set. To this end, the same two weighting schemes as before were used (see Section 2.2.1), resulting in the block-based global loss function (4) and the entry-based global loss function (5).

To study the recovery performance of both weighting schemes, again a distinction is drawn between recovery at the level of the model matrices on the one hand and recovery at the level of the component matrices on the other hand. Also, the same three types of $I \times J \times K$ and $I \times L$ block couples (as in Section 2.2.1) are distinguished, with these block couples now being binary instead of real-valued: ($\underline{\mathbf{T}}^1$, $\mathbf{T}^2$), ($\underline{\mathbf{D}}^1$, $\mathbf{D}^2$), and ($\underline{\mathbf{M}}^1$, $\mathbf{M}^2$).

The design and the procedure of the simulation study are outlined in Section 3.2.2. The simulation results are presented in Section 3.2.3.

3.2.2. Design and procedure

The same three factors as before were systematically manipulated in a completely randomized three-factorial design (all factors considered random): (1) the Array/total ratio, r, (2) the Dominance in size, d, and (3) the Error level, ε. While the

Array/total ratio and the Dominance in size were defined and manipulated in the same way as in the multiway multiblock component model (see Section 2.2.2), the *Error level*, ε, was defined as the expected proportion of cells of $(\underline{\mathbf{D}}^1, \mathbf{D}^2)$ differing from the corresponding cells of $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$:

$$\varepsilon = E\left[\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(t_{ijk}^1 - d_{ijk}^1)^2 + \sum_{i=1}^{I}\sum_{l=1}^{L}(t_{il}^2 - d_{il}^2)^2}{IJK + IL}\right]. \tag{14}$$

This factor was manipulated at four levels: .00, .10, .20, .30.

For each combination of the levels of the three factors, 20 (different replications of) true block couples $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ were constructed. In particular, each $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ couple was constructed as follows: Rank 4 true component matrices $\mathbf{A}^{(T)}$, $\mathbf{B}^{1^{(T)}}$, $\mathbf{C}^{(T)}$, and $\mathbf{B}^{2^{(T)}}$ were generated by independently sampling entries from a Bernoulli distribution with a parameter value of .50, subject to the constraint that all component-specific classes (i.e., a group of elements belonging to one component only) were non-empty. This constraint was imposed to ensure that: (1) the true block couple $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$, obtained by combining the true component matrices $\mathbf{A}^{(T)}$, $\mathbf{B}^{1^{(T)}}$, $\mathbf{C}^{(T)}$, and $\mathbf{B}^{2^{(T)}}$ by the INDCLAS and the HICLAS decomposition rules (11) and (12), could not be perfectly represented by a CHIC model of a lower rank than the True rank, which always equals 4 here, and (2) that the CHIC decomposition of $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ is unique (upon a permutation of the components). Note that the True rank of the CHIC model for $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ again was not manipulated here, because in a pilot study it turned out that this factor has no influence on the results. Next, a data block couple $(\underline{\mathbf{D}}^1, \mathbf{D}^2)$ was constructed for each true block couple $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ by changing the value in each cell of $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ with a probability ε. In total, 960 different couples $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ and $(\underline{\mathbf{D}}^1, \mathbf{D}^2)$ were obtained, i.e., 20 (replications) $\times$ 4 (Array/total ratio) $\times$ 3 (Dominance) $\times$ 4 (Error level). Finally, each data block couple $(\underline{\mathbf{D}}^1, \mathbf{D}^2)$ was subjected to two rank 4 CHIC analyses, one with the block-based loss function and one with the entry-based loss function. In order to lower the risk of ending in a local optimum, again for each analysis 50 runs of the CHIC algorithm (all with a random start) were performed and the solution with the lowest value on the loss function was retained.

3.2.3. Results

To study the recovery performance of both weighting schemes, again a distinction is drawn between the model matrices on the one hand and the (individual) component matrices on the other hand. With respect to the model matrices, the extent to which the entries of the true block couple are recovered, is quantified by computing the mean sum of squared differences between the model block couple $(\underline{\mathbf{M}}^1, \mathbf{M}^2)$ and the true block couple $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$. Further, with respect to the individual component matrices, the recovery performance was measured by computing for each component matrix a *goodness-of-recovery* statistic (*GOR*). For the common component matrix $\mathbf{A}$, this statistic was defined as follows:

$$1 - \frac{\sum_{i=1}^{I}\sum_{p=1}^{P}(a_{ip}^{(T)} - a_{ip}^{(M)})^2}{IP}, \tag{15}$$

with $a_{ip}^{(T)}$ and $a_{ip}^{(M)}$ denoting the (binary) elements of the true component matrix $\mathbf{A}^{(T)}$ and the obtained component matrix $\mathbf{A}^{(M)}$, respectively, and P denoting the rank of the model, which equals 4 here. Note that this *GOR* statistic always yields a value between 1, indicating perfect recovery, and 0, indicating no recovery at all. For the other component matrices, a similar *GOR* statistic was defined. Further, a combined *GOR* statistic (*cGOR*) was computed by a weighted average of the *GOR* for the object, attribute, source, and covariate component matrix, respectively, with the weights being proportional to the number of elements in the respective component matrices. Note that the (joint) permutational freedom of the components in $\mathbf{A}^{(M)}$, $\mathbf{B}^{1^{(M)}}$, $\mathbf{C}^{(M)}$, and $\mathbf{B}^{2^{(M)}}$ was accounted for by taking that permutation that maximizes the following criterion k:

$$k = \left[1 - \frac{\sum_{i=1}^{I}\sum_{p=1}^{P}(a_{ip}^{(T)} - a_{ip}^{(M)})^2}{IP}\right] + \left[1 - \frac{\sum_{j=1}^{J}\sum_{p=1}^{P}(b_{jp}^{1^{(T)}} - b_{jp}^{1^{(M)}})^2}{JP}\right]$$
$$+ \left[1 - \frac{\sum_{k=1}^{K}\sum_{p=1}^{P}(c_{kp}^{(T)} - c_{kp}^{(M)})^2}{KP}\right] + \left[1 - \frac{\sum_{l=1}^{L}\sum_{p=1}^{P}(b_{lp}^{2^{(T)}} - b_{lp}^{2^{(M)}})^2}{LP}\right]. \tag{16}$$

In Table 3, for both weighting schemes, the mean sum of squared differences between $(\underline{\mathbf{M}}^1, \mathbf{M}^2)$ and $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ is displayed, for the global model, consisting of the two submodels, as a whole (i.e., that again is obtained by taking all elements of both

Table 3
Mean recovery of true block couple entries for analyses with block-based and entry-based weighting schemes, calculated for the global model as a whole and for the separate submodels.

	Weighting scheme	
	Block-based	Entry-based
Global model	0.02157	0.00684
Submodels		
HICLAS	0.05551	0.03741
INDCLAS	0.01478	0.00013

Table 4
Mean recovery of true component matrices (*GOR*) for analyses with block-based and entry-based weighting schemes, calculated as a whole and for the separate component matrices.

	Weighting scheme	
	Block-based	Entry-based
Whole model (*cGOR*)	0.9518	0.9636
Separate component matrices		
A	0.9732	0.9996
$\mathbf{B}^1$	0.9972	0.9995
C	0.9974	0.9997
$\mathbf{B}^2$	0.9211	0.9310

blocks together in one superblock) as well as for the separate submodels. Just as was the case for the multiway multiblock component model, it turns out that, on average, the entries of the true block couple are better recovered when using the entry-based loss function than when using the block-based loss function, with the effect being largest for the two-way submodel (i.e., the HICLAS model).

In Table 4, the mean *GOR* value, in terms of the combined measure (*cGOR*) and the *GOR* measures for the separate component matrices, is displayed for both weighting schemes. Just as for the multiway multiblock component model, the entry-based loss function, on average, outperforms the block-based loss function. Note that this pattern is also observed for the individual component matrices, with the difference in recovery performance being largest for the common component matrix **A**.

The recovery performance of both weighting schemes was further compared by computing the following $cGOR^{diff}$ statistic:

$$cGOR^{diff} = cGOR^{entry} - cGOR^{block}. \tag{17}$$

Note that also an analogous *diff* statistic for the sum of squared differences between $(\underline{\mathbf{M}}^1, \mathbf{M}^2)$ and $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ could be computed, but that again only results for the $cGOR^{diff}$ statistic will be reported, because both statistics yield similar results. To investigate the relation between the difference in recovery performance between both weighting schemes on the one hand and the manipulated data characteristics on the other hand, in Fig. 2, the mean $cGOR^{diff}$ value is displayed as a function of the Array/total ratio, the Error level, and the Dominance in size. In an analysis of variance with $cGOR^{diff}$ as the dependent variable, it turned out that all main and interaction effects of the manipulated characteristics are significant at .05 level. Only reporting on effects with $\hat{\rho}_I \geq .10$, this analysis of variance further showed that the entry-based loss function especially outperforms the block-based loss function when the Error level increases ($\hat{\rho}_I = .10$) and the Array/total ratio increases ($\hat{\rho}_I = .16$). Both main effects, however, are qualified by a strong Error level by Array/total ratio interaction ($\hat{\rho}_I = .46$): The difference between the two weighting schemes is more pronounced when both the amount of error and the Array/total ratio are large. Note again that when $r = .50$ both weighting schemes perform equally well, because they coincide.

4. Discussion

In this paper, it was studied whether the data entries or the data blocks of a coupled data set should be considered as the units of information in order to best disclose the true structure underlying the data in case the data blocks differ considerably in size. To this end, two weighting schemes were compared in terms of their capability to recover the (common) underlying parameters of a coupled data set. This comparison was performed within the context of two global models for a coupled data set consisting of a three-way three-mode data block and a two-way two-mode data block that have one mode in common. In both cases, it appeared that the entry-based weighting scheme (i.e., taking the data entries as the units of information) outperformed the block-based weighting scheme (i.e., considering the data blocks as the units of information) in terms of recovery of the entries of (1) the true block couple, and (2) the true component matrices, with the largest difference in recovery performance being observed for the common component matrix **A**. Moreover, in both cases the difference in recovery performance appeared to increase when the data contain more error and when the three-way data block becomes larger relative to the two-way data block.

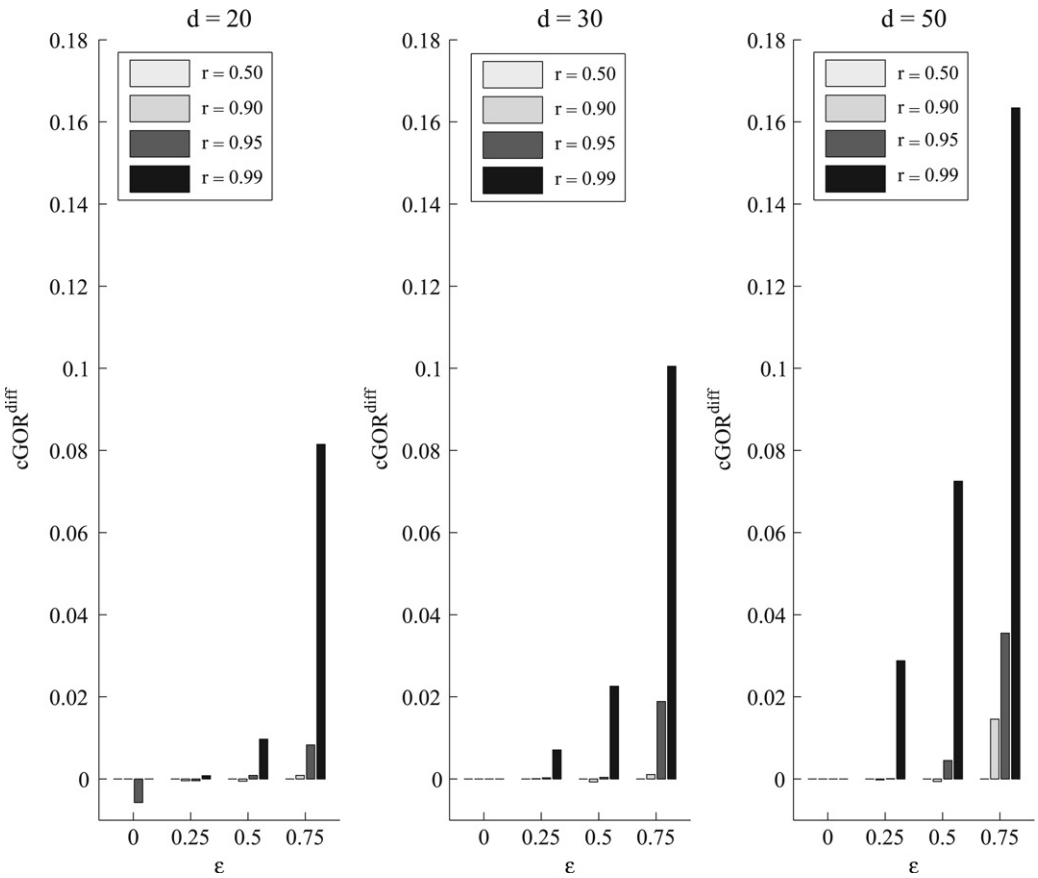

Fig. 2. Mean $cGOR^{diff}$ value as a function of the Array/total ratio (r), the Error level (ε), and the Dominance in size (d).

When considering these results, the question arises why the entry-based weighting scheme better uncovers the underlying structure of the coupled data than the block-based weighting scheme. In the remainder, two points of view will be outlined from which we try to answer this question.

As to the first point of view, for both global models it is observed, both at the level of the entries of the true block couple (see Tables 1 and 3) and at the level of the true component matrices (see Tables 2 and 4), that the recovery performance of the two-way submodel (i.e., the PCA and HICLAS submodel) is worse than the recovery performance of the three-way submodel (i.e., the CANDECOMP/PARAFAC and INDCLAS submodel), with this effect being most pronounced for the block-based weighting scheme. A possible explanation for these results could be that with regard to the two-way data the algorithm has a tendency to be misguided by the data at the expense of the truth, with this tendency becoming stronger when the two-way part is more emphasized, as is the case for the block-based weighting scheme.

As to the second point of view, considering that data entries contain noise, and assuming that the amount of noise is the same for all data entries, one might be tempted to conclude that it may be preferable to have each data entry influence the analysis to the same extent, which is equivalent to an entry-based weighting scheme. At this point, it is important to note that such a reasoning is based on an implicit stochastic framework. Stated differently, one may wonder whether an entry-based weighting scheme could be justified on the basis of implicit stochastic extensions of the deterministic multiway multiblock component model and the CHIC model.

In the remainder these two points of view will be discussed more in detail.

4.1. Comparing badness-of-fit to badness-of-recovery

To support the conjecture that the two-way part of both global models is more prone to be misguided by the data at the expense of the truth than the three-way part and that this is especially the case when the block-based weighting scheme is used, for each weighting scheme the ratio of the badness-of-recovery to the badness-of-fit is calculated, and this at the level of the separate submodels of the global model. As before, badness-of-recovery is defined as the sum of squared differences between the model block and the true block, while badness-of-fit is defined as the sum of squared differences between the model block and the data block, with both sums of squared differences being normalized by the size of the corresponding block. When the model would be pulled towards the truth, the badness-of-recovery will be small at the expense of the

Table 5
Mean ratio of badness-of-recovery to badness-of-fit for analyses with block-based and entry-based weighting schemes for both submodels of the multiway multiblock component model.

	Weighting scheme	
	Block-based	Entry-based
PCA	0.5538	0.1781
CANDECOMP/PARAFAC	0.0795	0.0142

Table 6
Mean ratio of badness-of-recovery to badness-of-fit for analyses with block-based and entry-based weighting schemes for both submodels of the CHIC model.

	Weighting scheme	
	Block-based	Entry-based
HICLAS	0.4482	0.2215
INDCLAS	0.0685	0.0009

badness-of-fit, implying a small ratio; when, however, the model would be misguided by the data at the expense of the truth, the opposite will be true, implying a large ratio. Note that in case of error-free data the ratio will necessarily equal one, because in that case the data block equals the true block and thus the badness-of-recovery equals the badness-of-fit.

Tables 5 and 6 display for the multiway multiblock component model and the CHIC model, respectively, the mean ratio of badness-of-recovery to badness-of-fit, computed at the level of the separate submodels. From these tables, it appears that the model is more misguided by the data at the expense of the truth for the two-way part of both global models than for the three-way part. Moreover, this phenomenon is more pronounced when the block-based weighting scheme is used than when the entry-based weighting scheme is used.

One can conclude that the block-based weighting scheme is outperformed by the entry-based weighting scheme because the former one gives relatively more weight to the two-way part of the global model, which has a (larger) tendency to be misguided by the data at the expense of the truth (than the three-way part), than the latter one. To elaborate on this conclusion, two cases are considered in which it may be expected that the recovery gain of the entry-based weighting scheme, in comparison to the block-based one, is lessened or even reversed. First, when the two-way data block becomes larger than the three-way data block, it could be expected that the entry-based weighting scheme is outperformed by the block-based one, because in that case the former scheme gives relatively more weight to the two-way part of the global model than the latter scheme. Second, when the model for the three-way part is replaced by a less restrictive model (i.e., a model that, given the same amount of data, has more free parameters to estimate), the three-way part will show a (stronger) tendency to be misguided by the data at the expense of the truth, implying that it may be expected that the difference in performance between both weighting schemes will decrease. In the following, the results of a new simulation study are presented in which, with respect to the multiway multiblock component model, for both cases, the performance of the two weighting schemes is evaluated.

As to the first case, a simulation study is performed in which the same design and procedure are used as explained in Section 2.2.2, but now with two other levels of the *Array/total ratio*: $r = .25$ and $r = .10$. Note that when generating these new data sets, the size of the three-way data block was kept constant (at 27 000 data elements), while the size of the two-way data block was increased in order to obtain $r = .25$ and $r = .10$. In Fig. 3, in which the mean $GOR_{\mathbf{A}}$ for the two weighting schemes is displayed as a function of the Array/total ratio, r (note the difference in scale between both parts of the figure), one can see that when the two-way data block is larger than the three-way data block, both weighting schemes yield almost perfect recovery, with the entry-based weighting scheme (still) outperforming the block-based one, although the difference is very small. The finding that the entry-based weighting scheme performs almost perfect even when the two-way data block is larger than the three-way data block may be clarified by taking into account how much information, in terms of number of data elements, the two-way data block contains. In particular, the two-way data block contains $I \times L$ data elements, with $I \times L > 27\,000$ when $r < .50$. As such, the tendency of the two-way part of the model to be misguided by the data at the expense of the truth is compensated by the large amount of information that is present in the two-way data block, and this mainly results in a better estimation of the common parameters. Note, however, that when $r < .50$, the two-way part of the model is still more misguided by the data than the three-way part, as is implied by the mean ratio of the badness-of-recovery to the badness-of-fit in these conditions being larger for the two-way part (i.e., 0.1649 and 0.1645 for the block- and entry-based weighting scheme in the $r = .10$ condition, respectively) than for the three-way part (i.e., 0.0108 and 0.0102 for the block- and entry-based weighting scheme in the $r = .10$ condition, respectively).

With respect to the second case, again the same design and procedure is used as explained in Section 2.2.2, but now with the rank 4 CANDECOMP/ PARAFAC model being replaced by a rank (4, 4, 4) Tucker3 model (Tucker, 1966). In order to test whether and how different numbers of components for the different modes and the structure of the core array (i.e., whether the core array has a simple structure in that only a few elements differ from zero or not, see Timmerman and Kiers (2000)) affect the choice of the weighting schemes, also data sets were generated and analyzed with a rank (4, 3, 5) Tucker3 model with a simple and a random core array. The results show that the difference in performance between

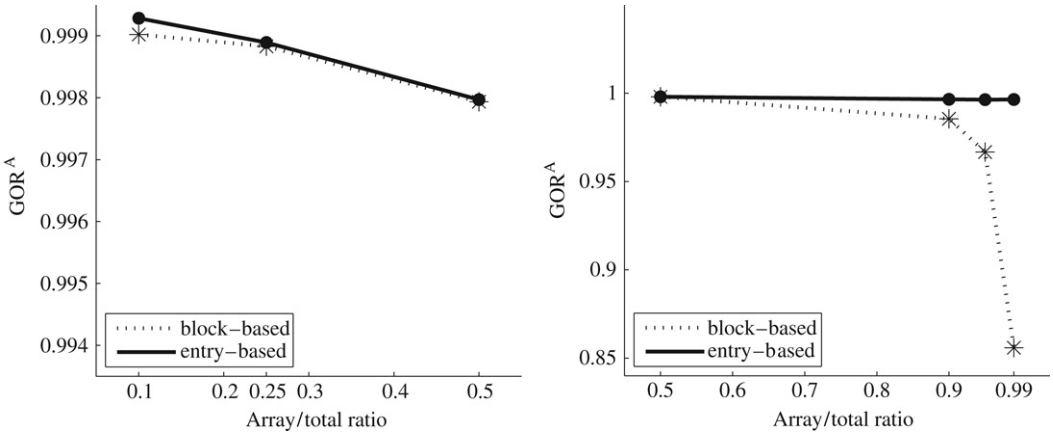

Fig. 3. Mean $GOR_{\mathbf{A}}$ for analyses with entry-based and block-based weighting schemes as a function of the Array/total ratio (r).

both weighting schemes hardly changes when the CANDECOMP/PARAFAC model is replaced by the less restrictive Tucker3 model. In particular, when the Tucker3 model is used, the mean sum of squared differences between $(\underline{\mathbf{T}}^1, \mathbf{T}^2)$ and $(\underline{\mathbf{M}}^1, \mathbf{M}^2)$ equals 0.00037 and 0.00016 for the block-based and entry-based weighting scheme, respectively (compared to 0.00039 and 0.00016 for the same weighting schemes when the CANDECOMP/PARAFAC model is used, see Table 1). At the level of the individual component matrices, when using the Tucker3 model, the mean *cGOR* equals 0.9503 and 0.9670 for the block-based and entry-based weighting schemes, respectively, while when using the CANDECOMP/PARAFAC model, the mean *cGOR* for both weighting schemes is equal to 0.9440 and 0.9643, respectively (see Table 2). Also for the different conditions of the design, replacing CANDECOMP/PARAFAC by Tucker3 does hardly influence the degree to which the entry-based weighting scheme outperforms the block-based one. Note that the difference in performance between both weighting schemes turns out not to be influenced by the rank of the Tucker3 model or the structure of the core array. These results may be explained by noting that when the three-way data block is not too small, the number of (extra) parameters that is added to the model by replacing CANDECOMP/PARAFAC by Tucker3 (i.e., the number of entries of the core array) is small relative to the number of parameters of the CANDECOMP/PARAFAC model. Note further that, because of the rotational freedom of the Tucker3 model (see e.g. Kiers (1992)), to fully identify the Tucker3 model the parameters of the model have to be restricted in some way, implying the loss of some free (extra) parameters (see e.g. Ramsay (1982)).

4.2. Post-hoc likelihood justification

Both for the multiway multiblock component model and for the CHIC model, a stochastic extension could be considered by postulating an appropriate distribution for the noise in the data (i.e., the deviations between the data and the model). Subsequently, the associated likelihood function of the data given the parameters could be computed. In particular, with regard to the multiway multiblock component model, one could assume that

$$d^1_{ijk} = m^1_{ijk} + e^1_{ijk}, \tag{18}$$

$$d^2_{il} = m^2_{il} + e^2_{il}, \tag{19}$$

with $e^1_{ijk} \sim N(0, \sigma^2)$, $e^2_{il} \sim N(0, \sigma^2)$, and the associated loglikelihood function being:

$$l(\underline{\mathbf{M}}^1, \mathbf{M}^2, \sigma) = \frac{IJK + IL}{2} \log(2\pi) + (IJK + IL) \times \log(\sigma) + \frac{1}{2\sigma^2}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(d^1_{ijk} - m^1_{ijk})^2 + \sum_{i=1}^{I}\sum_{l=1}^{L}(d^2_{il} - m^2_{il})^2\right]. \tag{20}$$

An analogous stochastic extension of the CHIC model again relies on (18), but now with $|e^1_{ijk}| \sim Bern(\pi)$ and $|e^2_{il}| \sim Bern(\pi)$, with π denoting the probability that a data entry differs from the corresponding model entry in both blocks (see also Leenen et al. (2008)). The associated loglikelihood function then reads:

$$l(\underline{\mathbf{M}}^1, \mathbf{M}^2, \pi) = \left[\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(d^1_{ijk} - m^1_{ijk})^2 + \sum_{i=1}^{I}\sum_{l=1}^{L}(d^2_{il} - m^2_{il})^2\right] \log\left[\frac{\pi}{1-\pi}\right] + (IJK + IL)\log(1-\pi). \tag{21}$$

When looking at the likelihood functions (20) and (21), it appears that optimizing these functions boils down to minimizing the sum across the two blocks of the sum of squared differences between the data and the model, which implies

that optimizing the likelihood is equivalent to minimizing the entry-based global loss function. The entry-based weighting scheme thus is optimal from a likelihood point of view. It has to be noted that for the presented stochastic extensions the parameter for the noise distribution (i.e., σ and π) is the same for both data blocks. One may wish to deal also with situations in which different data blocks are prone to different amounts of noise. Further research is needed to find out which weighting scheme performs best in such a situation.

4.3. Conclusion

When a two-way data block is coupled to a three-way data block by one common mode, and when both data blocks differ (considerably) in size but are subjected to the same amount of noise, a weighting scheme in which each data entry has the same amount of influence on the analysis better uncovers the true underlying structure of the coupled data, than a weighting scheme in which each data block influences the analysis to the same extent. This superior performance of the entry-based weighting scheme over the block-based one can be attributed to the fact that the former scheme gives relatively more weight to the data block that contains the most information. Also from a likelihood point of view an entry-based weighting scheme should be preferred over a block-based one.

Acknowledgements

The first author is a Research Assistant of the Fund for Scientific Research - Flanders (Belgium). The research reported in this paper was partially supported by the Research Council of K.U.Leuven (GOA/2005/04 and EF/2005/07, 'SymBioSys') and by IWT-Flanders (SBO 60045, 'Bioframe'). Requests for reprints should be sent to Tom Wilderjans, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium. Email: Tom.Wilderjans@psy.kuleuven.be.

References

Carroll, J.D., Chang, J.J., 1970. Analysis of individual differences in multidimensional scaling via an *n*-way generalization of "eckart-young" decomposition. Psychometrika 35, 283–319.
De Boeck, P., Rosenberg, S., 1988. Hierarchical classes: Model and data analysis. Psychometrika 53, 361–381.
Haggard, E.A., 1958. Intraclass Correlation and the Analysis of Variance. Dryden, New York.
Harshman, R.A., 1970. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis. UCLA Working Papers in Phonetics 16, pp. 1–84.
Kiers, H.A.L., 1992. Tuckals core rotations and constrained Tuckals modelling. Statistica Applicata 4, 659–667.
Kiers, H.A.L., 2000. Towards a standardized notation and terminology in multiway analysis. J. Chemometr. 14, 105–122.
Kirk, R.E., 1982. Experimental Design: Procedures for the Behavioral Sciences, 2nd edition. Brooks/Cole, Belmont, CA.
Kroonenberg, P.M., De Leeuw, J., 1980. Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika 45, 69–97.
Leenen, I., Van Mechelen, I., De Boeck, P., Rosenberg, S., 1999. INDCLAS: A three-way hierarchical classes model. Psychometrika 64, 9–24.
Leenen, I., Van Mechelen, I., Gelman, A., De Knop, S., 2008. Bayesian hierarchical classes analysis. Psychometrika 73, 39–64.
Ramsay, J.O., 1982. Some statistical approaches to multidimensional scaling data. J. Roy. Stat. Soc. A Stat. 145, 285–312.
Smilde, A.K., Kiers, H.A.L., 1999. Multiway covariates regression models. J. Chemometr. 13, 31–48.
Smilde, A.K., Westerhuis, J.A., Boqué, R., 2000. Multiway multiblock component and covariates regression models. J. Chemometr. 14, 301–331.
Timmerman, M.E., Kiers, H.A.L., 2000. Three-mode principal component analysis: Choosing the numbers of components and sensitivity to local optima. British. J. Math. Statist. Psych. 53, 1–16.
Tucker, L.R., 1951. A method for synthesis of factor analysis studies. Personnel Research Section Rapport # 984, Washington, DC: Department of the Army.
Tucker, L.R., 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31, 279–311.
Van Mechelen, I., Schepers, J., 2007. A unifying model involving a categorical and/or dimensional reduction for multimode data. Comput. Statist. Data Anal. 52, 537–549.
Wilderjans, T., Ceulemans, E., Van Mechelen, I., The chic -model: A global model for coupled binary data. Psychometrika. in press (doi:10.1007/s11336-008-9069-9).