Open access • Journal Article • DOI:10.1348/000711005X64817

Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. — Source link

Eva Ceulemans, Henk A.L. Kiers

Institutions: Katholieke Universiteit Leuven, University of Groningen

Published on: 01 May 2006 - British Journal of Mathematical and Statistical Psychology (John Wiley & Sons, Ltd)

Topics: Convex hull, Model selection, Heuristic and Principal component analysis

British Journal of Mathematical and Statistical Psychology (2006), 59, 133–150

The British Psychological Society
www.bpsjournals.co.uk

Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method

Eva Ceulemans[1]* and Henk A. L. Kiers[2]
[1]Katholieke Universiteit Leuven, Belgium
[2]Rijksuniversiteit Groningen, The Netherlands

Several three-mode principal component models can be considered for the modelling of three-way, three-mode data, including the Candecomp/Parafac, Tucker3, Tucker2, and Tucker1 models. The following question then may be raised: given a specific data set, which of these models should be selected, and at what complexity (i.e. with how many components)? We address this question by proposing a numerical model selection heuristic based on a convex hull. Simulation results show that this heuristic performs almost perfectly, except for Tucker3 data arrays with at least one small mode and a relatively large amount of error.

1. Introduction

The family of three-mode principal component models is a collection of methods for analysing three-way, three-mode data, for instance, scores of a number of participants on a number of variables, measured in a number of conditions. In particular, the family of three-mode principal component models consists of the Candecomp/Parafac (CP), Tucker3 (T3), Tucker2 (T2), and Tucker1 (T1) models (Carroll & Chang, 1970; Harshman, 1970; Kroonenberg, 1983; Kroonenberg & De Leeuw, 1980; Tucker, 1966). Being generalizations of standard two-mode principal component analysis (PCA), these models summarize the main information in the data by reducing up to three modes of the data to a few components and defining a linking structure between the components of the reduced modes and, if applicable, the elements of the modes not reduced. The formal relations among (most of) these models have been discussed by Kiers (1991) and Kroonenberg (1983).

*Correspondence should be addressed to Eva Ceulemans, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium (e-mail: eva.ceulemans@psy.kuleuven.be).

DOI:10.1348/000711005X64817

Considering that several models have been developed for the same type of data, the question may be raised as to which type of three-mode principal component model at what complexity (i.e. with how many components) yields the most useful description of a given data set. Hitherto, this complex model selection problem has almost always been addressed by first choosing a particular model type on the basis of substantive arguments and then selecting among solutions of the chosen model type with the aid of numerical model selection heuristics developed for the model type in question – for example, the CONCORDIA method for selecting among CP solutions of different complexities (Bro & Kiers, 2003) and the DIFFIT method for solving the T3 model selection problem (Kiers & der Kinderen, 2003; Timmerman & Kiers, 2000).

A more systematic approach for solving the three-mode principal component model selection problem may be the use of the visual inspection based method proposed by Kroonenberg and Van der Voort (1987), Kroonenberg and Oort (2003), and Murakami and Kroonenberg (2003). Following the seminal work of Mallows (1973), Verbeek (1984), and Fowlkes, Freeny, and Landwehr (1988), these authors have suggested handling similarly complex model selection problems by visually inspecting scree-like plots with a measure of the badness of fit (e.g. residual sum of squares) of the different solutions on the *y*-axis and a measure of the degrees of freedom associated with each solution on the *x*-axis. In particular, they argue that one should select a model on or close to an elbow in the lower boundary of the convex hull of this scree-like plot, because these 'hull' solutions have the best badness-of-fit/degrees-of-freedom balance.

In the present paper, building on the visual inspection based method, we propose to solve the three-mode principal component model selection problem by means of a numerical heuristic. In particular, we propose and evaluate a numerical procedure for assessing the boundary of the convex hull in scree-like plots as well as the elbow in the boundary. In contrast to the visual inspection based procedure, this numerical model selection heuristic can be programmed, which is important for two reasons. First, in practice this will help people with the often difficult and subjective task of choosing a model (although it should be noted that subjective aspects in this choice will and should always remain). Second, it allows for a systematic test of the validity of the numerical model selection procedure: applying a fully programmed model selection heuristic to data constructed in a simulation study, we can assess how often the procedure actually indicates the underlying model correctly. Such a simulation study will be reported in this paper. Furthermore, our numerical procedure differs slightly from the visual inspection method in two respects. First, we use goodness-of-fit rather than badness-of-fit measures, with, in some cases, these goodness-of-fit measures being approximate; note that the use of goodness-of-fit measures implies that we are interested in the higher rather than the lower boundary of the convex hull. Second, we use numbers of free parameters rather than degrees of freedom; note that Weesie and Van Houwelingen (1983) pointed out that for the T3 model the degrees of freedom equal the number of observations in the data minus the number of free parameters.

The rest of this paper is organized as follows. Section 2 describes the family of three-mode principal component models. In Section 3 the numerical convex hull based model selection heuristic is proposed. In particular, it is explained in detail how one may obtain approximate goodness-of-fit measures for some of the models and how the number of free parameters is defined. Furthermore, a description is given of how to find the higher boundary of the convex hull as well as an elbow in it. In Section 4 the proposed model selection heuristic is illustrated by applying it to an empirical data set. In Section 5 the performance of this heuristic is evaluated in an extensive simulation

study. Section 6 contains a theoretical and empirical comparison of the numerical convex hull based heuristic and Timmerman and Kiers's (2000) DIFFIT method for solving the T3 model selection problem. Section 7 contains some concluding remarks.

2. The family of three-mode principal component models

2.1. Models

Each three-mode principal component model approximates an $I \times J \times K$ participants by variables by conditions data array $\underline{\mathbf{X}}$ by a model array $\underline{\mathbf{M}}$ of the same size. Each model further includes a decomposition of $\underline{\mathbf{M}}$ into (a) up to three component matrices $\mathbf{A}$ ($I \times P$), $\mathbf{B}$ ($J \times Q$), and $\mathbf{C}$ ($K \times R$) that respectively reduce the participants, variables, and conditions to P, Q, and R components, and (b) a three-way, three-mode core array $\underline{\mathbf{G}}$ that defines a linking structure among the components of the reduced modes and, if applicable, the elements of the modes not reduced. In this paper, eight different types of three-mode principal component model are considered, which are all represented in Figure 1. In the following paragraphs we will discuss their distinctive features.

2.1.1. The three types of T1 model

A T1 model reduces only one of the three modes of $\underline{\mathbf{M}}$ to components. Hence, three different types of T1 model can be distinguished: type A which reduces the participants, type B which reduces the variables, and type C which reduces the conditions. Formally, the decomposition rules of the T1A, T1B, and T1C models can be stated as follows:

$$\text{T1A} : m_{ijk} = \sum_{p=1}^{P} a_{ip} g_{pjk}, \tag{1}$$

$$\text{T1B} : m_{ijk} = \sum_{q=1}^{Q} b_{jq} g_{iqk}, \tag{2}$$

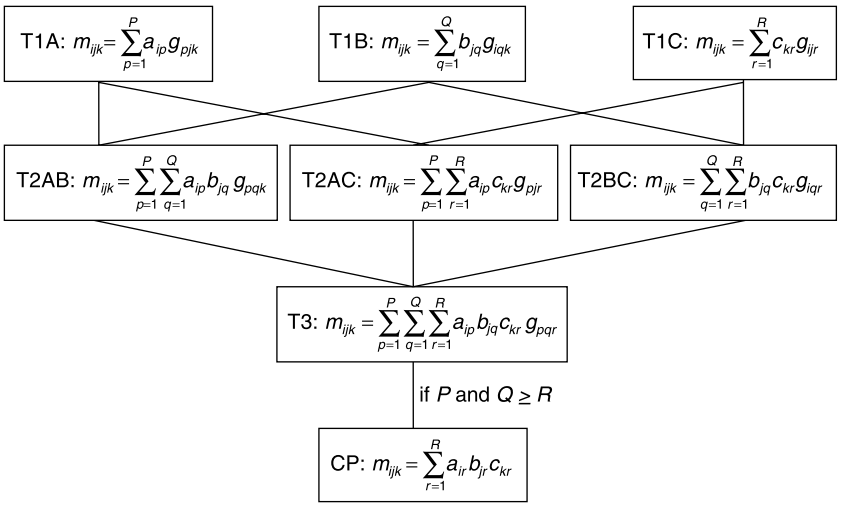

Figure 1. Interrelations of the eight types of three-mode principal component model.

and

$$\text{T1C} : m_{ijk} = \sum_{r=1}^{R} c_{kr} g_{ijr}, \tag{3}$$

with P, Q, and R indicating the complexity, that is, the number of components of the respective models. Note that a T1A model of complexity P is equivalent to a PCA model with P components for the matricized $I \times JK$ model array. The same holds for a T1B model of complexity Q (or a T1C model of complexity R) in being equivalent to a PCA model with Q (or R) components for the matricized $J \times KI$ (or $K \times IJ$) model array.

2.1.2. The three types of T2 model

A T2 model reduces two of the three modes of $\underline{\mathbf{M}}$ to components, implying that three different types of T2 model can be considered: type AB which reduces the participants and the variables, type AC which reduces the participants and the conditions, and type BC which reduces the variables and the conditions. The decomposition rules of these T2AB, T2AC, and T2BC models are given by

$$\text{T2AB} : m_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} a_{ip} b_{jq} g_{pqk}, \tag{4}$$

$$\text{T2AC} : m_{ijk} = \sum_{p=1}^{P} \sum_{r=1}^{R} a_{ip} c_{kr} g_{pjr}, \tag{5}$$

and

$$\text{T2BC} : m_{ijk} = \sum_{q=1}^{Q} \sum_{r=1}^{R} b_{jq} c_{kr} g_{iqr}, \tag{6}$$

with (P, Q), (P, R), and (Q, R) representing the complexity of the respective models.

2.1.3. The T3 model

A T3 model reduces each of the three modes of $\underline{\mathbf{M}}$ to components. Formally, the T3 decomposition rule reads

$$m_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr}, \tag{7}$$

with (P, Q, R) denoting the complexity of the model.

2.1.4. The CP model

A CP model summarizes each of the three modes of $\underline{\mathbf{M}}$ by the same number of components and restricts the core array $\underline{\mathbf{G}}$ to a unit superdiagonal array (i.e. $g_{pqr} = 1$ if and only if $p = q = r$), implying a one-to-one correspondence among the respective components. Hence, the CP decomposition rule can be stated as follows:

$$m_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}, \tag{8}$$

where R denotes the complexity of the model.

From the model descriptions above, one may conclude that some of the three-mode principal component models are interrelated in that some models are more restrictive or constrained versions of other models. All such interrelations among the models are graphically represented in Figure 1, which is to be read as follows. Model 1 is less restrictive than or as restrictive as model 2 if and only if a downward path of lines exists from model 1 to model 2. As such, one may, for instance, conclude that a T3 model of complexity (P, Q, R) is a constrained version of, amongst others, a T2BC model of complexity (Q, R) and a T1A model of complexity P. Similarly, one may conclude that a T1A model of complexity P and a T2BC model of complexity (Q, R) are not interrelated. For a more detailed description of the interrelations among the models, refer to Kiers (1991) and Kroonenberg (1983).

2.2. Fitting the models to data

To fit the eight types of three-mode principal component model to a given data array $\underline{\mathbf{X}}$, one may apply the standard two-mode PCA approach to the appropriately matricized data array for obtaining T1 solutions, and use the alternating least squares T2, T3, and CP algorithms for obtaining T2, T3, and CP solutions (for details of the T2 and T3 algorithms, see Kroonenberg & De Leeuw, 1980; for details about the CP algorithm, see Carroll & Chang, 1970; Harshman, 1970). Given a specific complexity and a data array $\underline{\mathbf{X}}$, these algorithms look for a model array $\underline{\mathbf{M}}$ that minimizes

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ijk} - m_{ijk})^2, \tag{9}$$

and that can be further decomposed according to a three-mode principal component model of the specified type and complexity. Subsequently, the goodness-of-fit value f of the obtained solution can be calculated as

$$f = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} m_{ijk}^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk}^2} \tag{10}$$

(Kroonenberg, 1983). Note that (10) holds irrespective of the preprocessing of the data.

3. A numerical convex hull based model selection heuristic

When applying each of the above three-mode PCA techniques to a given data array $\underline{\mathbf{X}}$, one typically faces a model selection problem: which of the eight different types of three-mode principal component model yields the most useful description of $\underline{\mathbf{X}}$, and at what complexity? The purpose of the present paper is to provide a solution for this complex problem by proposing a numerical convex hull based model selection procedure. As mentioned in the Introduction, this procedure numerically assesses the higher boundary of the convex hull of a plot of goodness of fit versus number of free parameters as well as the location of the elbow in this boundary. In this section, we present the details of the proposed procedure. We begin by describing how to obtain

approximate goodness-of-fit measures for some of the models. Then we examine how the number of free parameters is defined for all models. This is followed by a description of the procedure for numerically assessing the higher boundary of the convex hull and the procedure for finding the elbow in this boundary. Finally, we give a stepwise overview of the proposed model selection heuristic.

3.1. Kiers and der Kinderen's quick method for obtaining approximate T3 and T2 goodness-of-fit values

Obtaining optimal T3 goodness-of-fit values with the alternating least squares T3 algorithm is rather time-consuming, especially if one wishes to consider several complexities. As this is an important factor for model selection procedures, Kiers and der Kinderen (2003) proposed a quick procedure for computing approximate T3 goodness-of-fit values for all possible complexities in one go, and demonstrated that applying Timmerman and Kiers's (2000) T3 model selection method DIFFIT to the *approximate* goodness-of-fit values yields slightly better results than applying DIFFIT to the optimal goodness-of-fit values (i.e. obtained with the alternating least squares T3 algorithm).

Kiers and der Kinderen's (2003) procedure works as follows. First, matricize $\underline{\mathbf{X}}$ into $\mathbf{X}_a\,(I \times JK)$, $\mathbf{X}_b\,(J \times KI)$, and $\mathbf{X}_c\,(K \times IJ)$ and compute the eigendecompositions $\mathbf{X}_a\mathbf{X}_a' = \mathbf{K}_a\mathbf{\Lambda}_a\mathbf{K}_a'$, $\mathbf{X}_b\mathbf{X}_b' = \mathbf{K}_b\mathbf{\Lambda}_b\mathbf{K}_b'$, and $\mathbf{X}_c\mathbf{X}_c' = \mathbf{K}_c\mathbf{\Lambda}_c\mathbf{K}_c'$. Then, compute $\mathbf{H}_a = \mathbf{K}_a'\mathbf{X}_a(\mathbf{K}_c \otimes \mathbf{K}_b)$, where $\mathbf{H}_a$, $\mathbf{K}_a$, $\mathbf{K}_b$, and $\mathbf{K}_c$ are the matricized version of the T3 core array $\underline{\mathbf{H}}$ and the T3 component matrices associated with the maximal number of components I, J, and K for the three modes, and $\otimes$ denotes the Kronecker product. Finally, compute the approximate goodness-of-fit value f of a T3 solution of complexity (P, Q, R) by dividing the sum of squared elements of the subarray $\underline{\mathbf{G}}$ of $\underline{\mathbf{H}}$ that is obtained by retaining only the first P participant, Q variable, and R condition components, by the sum of squared elements of $\underline{\mathbf{X}}$:

$$f = \frac{\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}^2}. \tag{11}$$

Replacing $\mathbf{K}_c$, $\mathbf{K}_b$, or $\mathbf{K}_a$ by an identity matrix, the same procedure can be used to obtain approximate goodness-of-fit values f of T2AB, T2AC, or T2BC solutions.

Finally, note that from the eigendecompositions of $\mathbf{X}_a$, $\mathbf{X}_b$, and $\mathbf{X}_c$, one can also calculate the optimal goodness-of-fit values f of T1A, T1B, and T1C solutions: divide the sum of squared elements of the submatrix of $\mathbf{\Lambda}_a$, $\mathbf{\Lambda}_b$, and $\mathbf{\Lambda}_c$ that is obtained by retaining only the first P participant, Q variable, and R condition components, respectively, by the sum of squared elements of $\underline{\mathbf{X}}$.

3.2. The number of free parameters of three-mode principal component solutions

Regarding the number of free parameters fp of the different types of three-mode principal component solution, Weesie and Van Houwelingen (1983) argue that the fp-value of a T3 solution of complexity (P, Q, R) amounts to $IP + JQ + KR + PQR - P^2 - Q^2 - R^2$. The last three terms of this sum correct for the fact that $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are determined up to non-singular transformations only. In other words, we may fix a $P \times P$ block of P^2 elements in $\mathbf{A}$, a $Q \times Q$ block in $\mathbf{B}$, and a $R \times R$ block in $\mathbf{C}$, because we can always (i.e. with probability 1) transform our solution without loss of fit into a solution with fixed elements thus.

By the same reasoning, the number of free parameters *fp* of a T2BC solution of complexity (Q, R) equals $JQ + KR + IQR - Q^2 - R^2$, with the last two terms correcting for transformational freedom. Note that the *fp*-value of a T3 model of complexity (I, Q, R), that is, a T3 model that is equivalent to a T2BC model of complexity (Q, R), gives exactly the same number of free parameters, as it should: $I^2 + JQ + KR + IQR - I^2 - Q^2 - R^2 = JQ + KR + IQR - Q^2 - R^2$. Regarding the equivalence relation between a T3 model and a T2BC model in general, note that a T3 model of complexity (I, Q, R) with $I \geq QR$ is equivalent to a T3 model of complexity (QR, Q, R) (Wansbeek & Verhees, 1989); hence, as a T2BC model of complexity (Q, R) is equivalent to a T3 model of complexity (I, Q, R), it is also equivalent to a T3 model of complexity (QR, Q, R). The T2AB and T2AC models have similar equivalence relations to the T3 model.

Again analogously, the number of free parameters *fp* of a T1A solution can be calculated as $IP + PJK - P^2$. Note again that the *fp*-value of a T1A solution of complexity P equals the *fp*-value of an equivalent T3 solution, that is, a T3 solution of complexity (P, J, K): $IP + J^2 + K^2 + PJK - P^2 - J^2 - K^2 = IP + PJK - P^2$.

Regarding the number of free parameters *fp* of a CP solution, given that CP solutions are determined up to scaling only, this number amounts to $(I + J + K)R - 2R$. Specifically, the last term corrects for the scaling freedom, where scaling the component matrices of two modes fixes the scaling of the third.

Finally, note that in cases where the size of one of the modes is larger than the product of the other two, special adjustments must be made, because in such cases (see Kiers & Harshman, 1997), the biggest mode can be reduced considerably without loss of information essential for the components of the two other modes and the core. Specifically, when $I > JK$, the data can be reduced to a $JK \times J \times K$ data set. Therefore, in such cases, in the computation of the number of free parameters *fp*, I should be replaced by JK.

3.3. Ceulemans and Van Mechelen's procedure for finding the solutions on the higher boundary of the convex hull

In this section, Ceulemans and Van Mechelen's (2005) procedure for determining the subset of solutions that are on the higher boundary of the convex hull of the plot of goodness of fit versus number of free parameters is described. The plots in Figure 2 will be used as a guiding example. Figure 2a shows a plot of goodness of fit versus number of free parameters for a data set from the simulation study in Section 5; in this plot the higher boundary of the convex hull has been drawn. As we wish to explain how to determine this boundary *numerically*, the 'clean' plot without the boundary is also given (Figure 2b).

Ceulemans and Van Mechelen's (2005) procedure works as follows. First, for each observed number of free parameters *fp*, retain only the best-fitting solution; if two or more solutions have equal *fp*- and *f*-values, select one of them at random. The solutions thus retained for our guiding example are displayed in Figure 2c. Considering the difference between Figures 2b and 2c, it can be concluded that this step implies a huge decrease in the number of solutions to be compared. Then, sort the n retained solutions by their number of free parameters *fp* and label them s_i $(i = 1, \ldots, n)$. In Figure 2c they are ordered from left to right. Next, given that the procedure looks for the solutions with an optimal balance between goodness of fit and number of free parameters, exclude a solution s_i from the n retained solutions if a solution s_j $(j < i)$ exists such that $f_j > f_i$; as can be seen from Figures 2c and 2d, this step implies that the line that one could draw between the points representing the subsequent solutions becomes

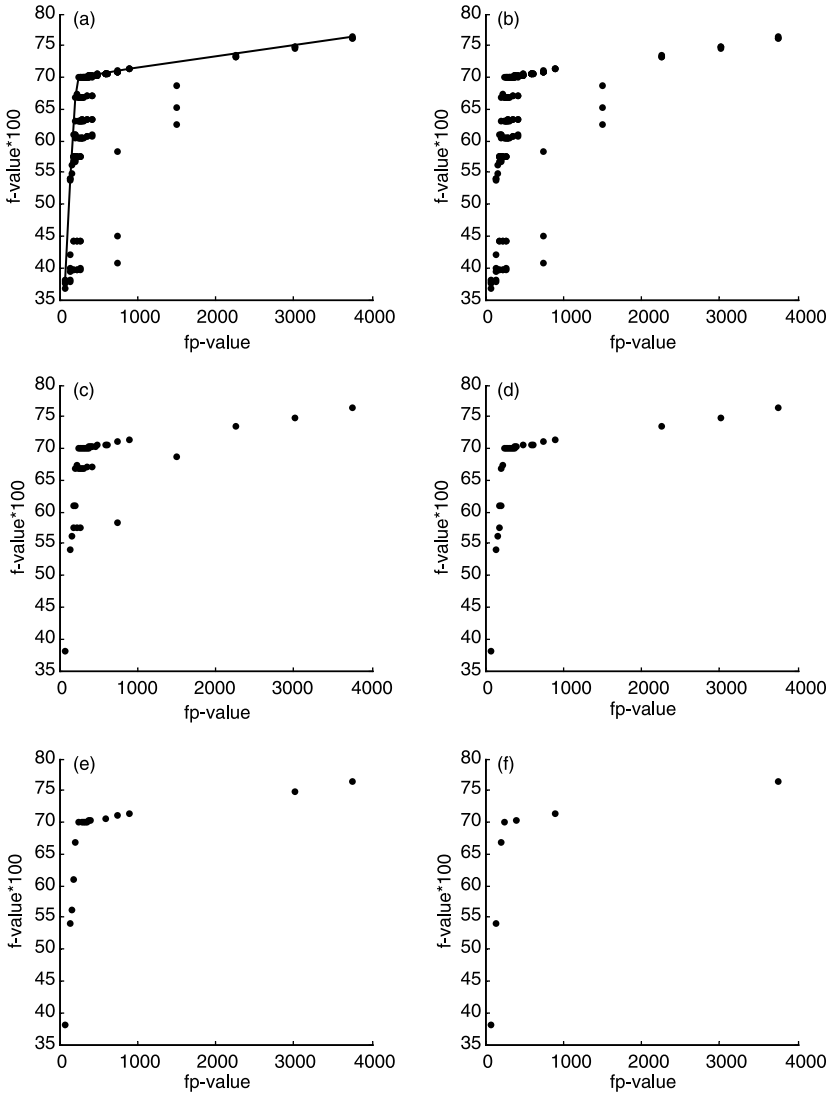

Figure 2. Graphical representation of the different steps (numbered as in Section 3.5) of Ceulemans and Van Mechelen's procedure for finding the solutions on the higher boundary of the convex hull of a plot of goodness of fit versus number of free parameters: (a) plot of goodness of fit versus number of free parameters, with the line representing the higher boundary of the convex hull; (b) retained solutions after step 1; (c) retained solutions after steps 2 and 3; (d) retained solutions after step 4; (e) retained solutions after step 5; and (f) retained solutions after step 6.

non-decreasing. Finally, apply the following routine that consecutively considers all triplets of adjacent solutions. For the first triplet of solutions (s_1, s_2, s_3), determine whether or not the point for s_2 is located below or on the line that one could draw between the points for s_1 and s_3 in the plot of goodness of fit versus number of free parameters; if so, exclude s_2 from the subset of retained solutions. Next, do the same for all following triplets of adjacent retained solutions; one run of this routine transforms Figure 2d into Figure 2e. Repeating this routine until no solution can be excluded yields

the solutions on the higher boundary of the convex hull of the plot of goodness of fit versus number of free parameters; Figure 2f, which shows the final set of retained solutions, does indeed consist of the solution points that belong to the higher boundary of the convex hull in Figure 2a.

3.4. Selecting among the hull solutions: A numerical implementation of the scree test

To select the solution on the higher boundary of the convex hull with the best balance of goodness of fit and number of free parameters *fp*, we propose the following numerical implementation of the scree test: select the solution i that maximizes

$$st_i = \frac{f_i - f_{i-1}}{fp_i - fp_{i-1}} \Big/ \frac{f_{i+1} - f_i}{fp_{i+1} - fp_i}. \tag{12}$$

A relatively large *st*-value indicates that allowing for fp_i free parameters (instead of fp_{i-1} free parameters) increases the fit of the model considerably, whereas allowing for more than fp_i free parameters hardly increases it at all. Thus, we select that solution after which the increase in fit levels off.

3.5. Stepwise overview of the numerical convex hull based model selection procedure

The numerical convex hull based model selection procedure can be summarized in eight steps:

(1) Determine the *fp*- and *f*-values of all three-mode principal component solutions from which one wishes to choose.
(2) For each of the n observed *fp*-values, retain only the best-fitting solution.
(3) Sort the n retained solutions by their *fp*-values and denote them by $s_i\,(i = 1, \ldots, n)$.
(4) Exclude all solutions s_i for which a solution $s_j\,(j < i)$ exists such that $f_j > f_i$.
(5) Consecutively consider all triplets of adjacent solutions. Exclude the middle solution if its point is located below or on the line connecting its neighbours in a plot of goodness of fit versus number of free parameters.
(6) Repeat step 5 until no solution can be excluded.
(7) Determine the *st*-values of the 'hull' solutions obtained.
(8) Select the solution with the highest *st*-value.

Note that equivalent solutions – for instance, a T3 model of complexity (*QR*, *Q*, *R*) and a T2BC model of complexity (*Q*, *R*) – have equal *fp*-values (see Section 3.2). Hence, the 'hull'-finding procedure described in Section 3.3 will retain at most one of such equivalent solutions, in particular, the solution with the highest *f*-value. However, some of these *f*-values are only approximate. Therefore, if one such equivalent solution has the highest *st*-value, one could choose to report one of the other equivalent solutions as well. In fact, the ultimate choice among them should then be made on the basis of ease of substantive interpretation.

Finally, in practice, the procedure can be used in a somewhat relaxed way: models close to the higher boundary could also be considered as important alternative models, as well as models with a high but not maximal scree test value *st*. Moreover, one should also take substantive considerations into account when selecting a model. However, the strict numerical procedure described above can be helpful as an important heuristic in

making a first selection. In the simulation study described in Section 5 it will be seen that this procedure works very well indeed.

4. Illustrative application

In this section we illustrate the use of the numerical convex hull based model selection heuristic by applying it to the Chopin's preludes data set, which can be downloaded from http://three-mode.leidenuniv.nl. As Murakami and Kroonenberg (2003) describe in detail, this data set was gathered by asking 38 Japanese university students to rate the 24 preludes composed by Chopin on 20 bipolar scales (e.g. bright–dark, slow–fast). Murakami and Kroonenberg suggested preprocessing the resulting $24 \times 20 \times 38$ data array $\underline{\mathbf{X}}$ by centring the scores across the prelude mode and then normalizing the scores by scale.

Given $\underline{\mathbf{X}}$, numbers of free parameters *fp* and goodness-of-fit values *f* were obtained for 169 three-mode principal component solutions: 15 T1A, T1B, and T1C solutions of complexity 1 to 5; 75 T2AB, T2AC, and T2BC solutions of complexity (1,1) to (5,5); 74 T3 solutions of complexity (1,1,1) to (5,5,5); and 5 CP solutions of complexity 1 to 5. With respect to the number of T3 solutions considered, note that, as mentioned in Section 3.2, Wansbeek and Verhees (1989) proved that a T3 model for which the number of components for the participant mode (for example) exceeds the product of the number of components for the other two modes gives the same fit as a T3 model for which the number of participant components equals the product of the number of components for the other two modes; this implies that of all 125 T3 solutions the 51 solutions for which $P > QR$, $Q > PR$, or $R > PQ$ can be omitted. The *fp*-values of the 169 solutions were computed as described in Section 3.2. The *f*-values of the T1A, T1B, and T1C solutions were calculated on the basis of the eigendecomposition of $\mathbf{X}_a$, $\mathbf{X}_b$, and $\mathbf{X}_c$, and the *f*-values of the T2AB, T2AC, T2BC, and T3 solutions were approximated with the Kiers and der Kinderen (2003) procedure (see Section 3.1). The *f*-values of the CP solutions resulted from analysing the data set with the CP algorithm (the best fit was retained from five runs, four of which were initialized randomly and one rationally).

Figure 3 shows a plot of goodness of fit versus number of free parameters for the 169 solutions. Applying the numerical convex hull based model selection procedure to this plot yielded 11 'hull' solutions, which are indicated by larger points in Figure 3. The scree test values *st* of these 11 'hull' solutions are given in Table 1. From this table, one may conclude that the 'hull' heuristic indicates the selection of the T2BC model of complexity (2,1). As this solution is equivalent to the T2AC solution of complexity (2,1) and the T3 solution of complexity (2,2,1), we could choose to report either of these two models as well. As mentioned in Section 3.5, the ultimate choice should be made on the basis of ease of substantive interpretation. Finally, it is interesting to note that Murakami and Kroonenberg (2003) reported that applying the DIFFIT method to optimal T3 solutions of complexity (1,1,1) to (3,3,3) yielded the same model selection result, in that it indicated the selection of the T3 solution of complexity (2,2,1).

5. Simulation study

In this section, we present an extensive simulation study in which we evaluate to what extent the numerical convex hull based model selection heuristic succeeds in indicating the type and complexity of the three-mode principal component model that underlies a given three-way, three-mode data array. In this simulation study, we distinguish between

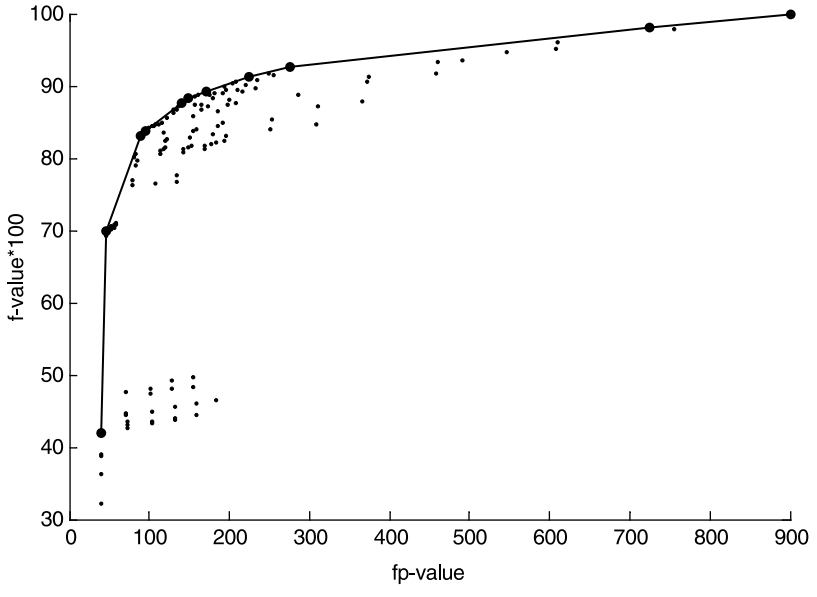

Figure 3. Plot of goodness of fit versus number of free parameters for the 169 three-mode principal component solutions for the Chopin's preludes data, with the line representing the higher boundary of the convex hull and the larger points indicating the 11 'hull' solutions.

two types of simulated data set: (a) data sets that are based on a randomly generated three-mode principal component model and (b) data sets that are constructed from an empirically obtained three-mode principal component solution.

5.1. Data sets constructed from a randomly generated three-mode principal component solution

5.1.1. Design

In this part of the simulation study, we constructed 225 data arrays for all eight types of three-mode principal component model. In particular, for each model type, three

Table 1. Goodness-of-fit values *f*, numbers of free parameters *fp*, and scree test values *st* of the 11 solutions on the higher boundary of the convex hull of Figure 3

Model	Complexity	*f*	*fp*	*st*
CP	1	.2893	80	–
T2BC	(2,1)	.4066	121	8.65
CP	2	.4195	160	1.01
CP	3	.4456	240	1.05
CP	4	.4705	320	1.28
CP	5	.4900	400	2.27
T2BC	(4,5)	.5232	709	1.25
T2AB	(5,4)	.5413	919	1.19
T2AB	(5,5)	.5558	1120	1.20
T1B	4	.7118	3712	1.60
T1B	5	.7466	4635	–

parameters were systematically varied:

(1) the size $I \times J \times K$ of the data arrays, at three levels: $200 \times 10 \times 10$, $50 \times 20 \times 20$, $27 \times 27 \times 27$;
(2) the true complexity of the three-mode principal component model that underlies the data arrays, at three levels 2, 3, 4 for T1A, T1B, T1C, and CP, at three levels (3,2), (3,3), (3,4) for T2AB, T2AC, and T2BC, and at three levels (3,2,2), (3,3,3), (4,3,2) for T3;
(3) the amount of error in the data, at five levels 0, 15, 30, 45, 60%.

For each cell of the design five replications were considered.

The 225 T3 data arrays (3 sizes × 3 complexities × 5 error levels × 5 replications) were generated by

$$\mathbf{X} = \mathbf{AG}(\mathbf{C}'\otimes\mathbf{B}') + \varepsilon\mathbf{E}, \tag{13}$$

where **A** is sampled from the standard normal distribution, **G** is sampled from a uniform distribution with entries ranging from -0.5 to 0.5, **B** and **C** are random orthonormal matrices, ε denotes a coefficient for manipulating the error level, and **E** is sampled from the standard normal distribution and multiplied by a scalar such that $\|\mathbf{AG}(\mathbf{C}'\otimes\mathbf{B}')\| = \|\mathbf{E}\|$. Note that $\mathbf{X}\,(I \times JK)$ and $\mathbf{G}\,(P \times QR)$ are matricized versions of $\underline{\mathbf{X}}$ and $\underline{\mathbf{G}}$. To generate the CP data arrays by means of (13), **A**, **B**, and **C** are all sampled from the standard normal distribution and **G** is fixed such that rewriting **G** into $\underline{\mathbf{G}}$ yields a unit superidentity array. To obtain T2 and T1 data arrays from (13), the component matrices of the reduced modes are random orthonormal matrices and the component matrices of the modes not reduced are fixed to identity matrices. This construction method may sometimes lead to data sets of which the structural part, $\mathbf{AG}(\mathbf{C}'\otimes\mathbf{B}')$, can be fitted almost as well by solutions with lower *fp*-values than the true solution. As Timmerman and Kiers (2000) argue that such cases may be undesirable for evaluating T3 model selection results, they only considered T3 data arrays for which at most 98% of the structural sum of squares can be fitted by T3 solutions with lower *fp*-values than the true solution. As we wanted to compare the performance of the 'hull' heuristic and DIFFIT (see Section 6), we did the same here for the T3 data. In order to keep the evaluation of the 'hull' heuristic as general as possible, however, we did not impose this constraint when generating CP, T2, and T1 data. Therefore, in the results section, we will investigate whether possible selection errors can be explained by this phenomenon.

For each of the 1800 simulated data sets (8 model types × 225 data arrays), numbers of free parameters *fp* and goodness-of-fit-values *f* were obtained for 565 three-mode principal component solutions (in the same way as for the example data set in Section 4): 24 T1A, T1B, and T1C solutions of complexity 1 to 8; 192 T2AB, T2AC, and T2BC solutions of complexity (1,1) to (8,8); 341 T3 solutions of complexity (1,1,1) to (8,8,8) (regarding the number of T3 solutions considered, refer to Section 4); and 8 CP solutions of complexity 1 to 8. The numerical convex hull based model selection heuristic was then applied to these 565 *fp*- and *f*-values.

5.1.2. Results

Given the 565 *fp*- and *f*-values, the numerical convex hull based model selection heuristic selected the correct model type and complexity for all 225 T1A, T1B, T1C, and T2BC data sets. For the 225 T2AB, T2AC, and CP data sets, correct selection occurred in all but one case for each model type; in the two T2 'incorrect selection' cases the

solution selected had a lower *fp*-value than the true solution but, when fitted to the structural data, accounted for more than 99.9% of the variance in the structural data. This indicates that, in fact, for these data sets the true and the selected model were virtually equivalent. For the CP 'incorrect selection' case, no such near equivalence was found.

For the 225 T3 data sets, an incorrect selection occurred in 17 cases. A further investigation of the results for these 17 cases shows that incorrect selection mostly occurs for size $200 \times 10 \times 10$, complexity (3,2,2) or (4,3,2), and 45 or 60% error (see Figure 4). In particular, in 10 of the 11 'incorrect selection' cases of complexity (3,2,2) the procedure selected the CP solution of complexity 2 and in 4 of the 5 cases of complexity (4,3,2) the T3 solution of complexity (3,3,2), with only one of these 17 T3 selection errors being explained by a virtual equivalence between the true and the selected model. To check whether the model selection problems for T3 data arrays of size $200 \times 10 \times 10$ are caused by (a) the asymmetry of these data arrays or (b) the small number of elements in the second and third modes, which yields a small number of second and third mode component entries in comparison to the number of core entries, we also generated and analysed 75 T3 data arrays of size $10 \times 10 \times 10$ and 75 T3 data arrays of size $625 \times 25 \times 25$; these data arrays were generated and analysed as described in Section 5.1.1. The results clearly show that the numerical convex hull based model selection heuristic does not perform so well for T3 data arrays for which at least one of the three modes contains few elements, say fewer than 15: whereas for size $625 \times 25 \times 25$ correct selection occurred in all but one case, an incorrect selection occurred for 32 of the 75 size $10 \times 10 \times 10$ data arrays. Also, given that for the T3 and T2 models approximate goodness-of-fit values were used whereas the CP and T1 goodness-of-fit values were optimal, an investigation was undertaken into whether (some of) the 17 T3 model selection problems could be solved by using optimal T3 goodness-of-fit values instead of approximate ones; this was not the case.

Finally, it should be noted that of the 20 T2AB, T2AC, CP, and T3 'incorrect selection' cases, the true solution did not belong to the higher boundary of the convex hull in two cases only. With respect to the 18 other cases, in 11 the true solution had the second highest *st*-value, in 4 it had the third highest *st*-value, and in 3 it had the fourth, fifth and eighth highest *st*-value, respectively. These results qualify the term 'incorrect selection', because they imply that if the numerical convex hull based procedure had been applied

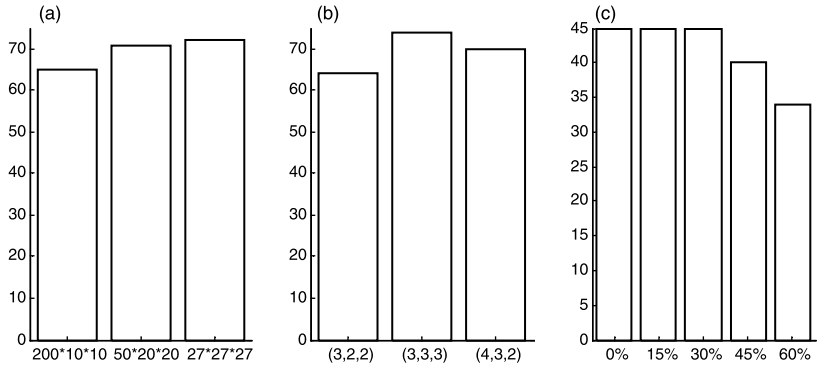

Figure 4. Frequency of correct model selection for the 225 T3 data arrays, as a function of (a) size, (b) complexity, and (c) error.

in a somewhat relaxed way (see Section 3.5), as will often be the case in practice, the true solution would in most cases have belonged to the subset of solutions selected for further consideration.

5.2. Data sets constructed from an empirically obtained solution

5.2.1. Design

In this part of the simulation study, we generated data arrays by adding error to the (2,2,1) T3 solution that was obtained for the Chopin's preludes data in Section 4. Similar to the construction of the randomly generated data sets, the error was sampled from the standard normal distribution and rescaled to obtain four levels of error perturbation: 15, 30, 45, and 60%. For each level of error five replications were considered.

For each of the 20 simulated data sets (4 error levels × 5 replications), numbers of free parameters *fp* and goodness-of-fit values *f* were obtained for 169 three-mode principal component solutions: 15 T1A, T1B, and T1C solutions of complexity 1 to 5; 75 T2AB, T2AC, and T2BC solutions of complexity (1,1) to (5,5); 74 T3 solutions of complexity (1,1,1) to (5,5,5) (regarding the number of T3 solutions considered, refer to Section 4); and 5 CP solutions of complexity 1 to 5. The numerical convex hull based model selection heuristic was then applied to the 169 *fp*- and *f*-values for these solutions.

5.2.2. Results

Applying the numerical convex hull based model selection heuristic to the 169 *fp*- and *f*-values for the 20 generated data sets always resulted in the selection of the true model, that is, the T3 solution of complexity (2,2,1) or the equivalent T2AC or T2BC solutions of complexity (2,1). However, when we used the T3 solution of complexity (3,2,2), which was reported by Murakami and Kroonenberg (2003), as a basis for generating simulated data sets, applying the 'hull' heuristic also resulted in half of the cases in the selection of the T3 solution of complexity (2,2,1) or the equivalent T2AC or T2BC solutions of complexity (2,1). This result is probably caused by the near equivalence of the true solution of complexity (3,2,2) and the selected solution of complexity (2,2,1): the solution selected accounts for 95.6% of the variance in the model array $\underline{\mathbf{M}}$ that is associated with the true solution.

6. Comparison between the proposed heuristic and Timmerman and Kiers's DIFFIT method

In this section, the numerical convex hull based model selection heuristic is compared to Timmerman and Kiers's (2000) DIFFIT method for selecting among T3 solutions of different complexities. DIFFIT works as follows:

(1) For all T3 solutions among which one wishes to select, determine the goodness-of-fit values *f* and the sum of components $sum = P + Q + R$.
(2) For each of the *N* observed *sum*-values, retain only the best-fitting solution. Denote the *N* retained solutions by s_{sum}.
(3) For each of the *N* retained solutions, compute dif_{sum} as the difference between $f_{s_{sum}}$ and $f_{s_{sum-1}}$; this implies that the *dif*-value of the simplest solution equals its goodness-of-fit value *f*.

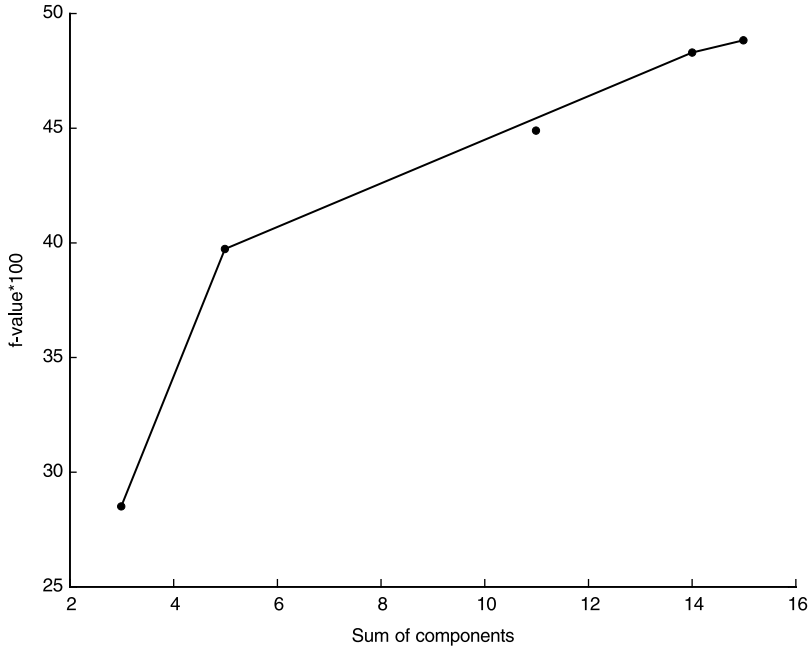

Figure 5. Plot of goodness of fit versus sum of components for the five T3 solutions for the Chopin's preludes data that are retained by the DIFFIT method, with the line representing the higher boundary of the convex hull.

(4) Exclude all solutions s_i for which a solution s_j $(j > i)$ exists such that $dif_j > dif_i$. Indicate the M remaining solutions by $m = 1, \ldots, M$. The associated *sum*-values are given by $sum(m)$, implying that the corresponding *dif*-values are given by $dif_{sum(m)}$. Note that this step results in an approximation of the set of solutions on the higher boundary of the convex hull of a plot of goodness of fit versus sum of components (see Kroonenberg & Oort, 2003). In particular, DIFFIT sometimes also retains solutions that are located close to but below the higher boundary of the convex hull. For an example see Figure 5, which displays a plot of goodness of fit versus sum of components for the retained solutions for the Chopin's preludes data set (see Section 4); it is clear that the retained solution with 11 components is located below the higher boundary of the convex hull.
(5) For the first $M - 1$ solutions, compute $b_{sum(m)} = dif_{sum(m)}/dif_{sum(m+1)}$.
(6) To eliminate solutions which entail small fit increases only, select those solutions for which $dif_{sum(m)} > \|\underline{\mathbf{X}}\|^2/(sum_{\max} - 3)$ only, with $sum_{\max} = \min(I, JK) + \min(J, IK) + \min(K, IJ)$.
(7) From the remaining solutions, select the solution with the highest $b_{sum(m)}$-value.

For a detailed discussion of the DIFFIT method refer to Timmerman and Kiers (2000).

6.1. Theoretical comparison

The above description suffices to note five differences between DIFFIT and the proposed numerical convex hull based model selection heuristic:

(1) Whereas the numerical convex hull based model selection heuristic uses the number of free parameters as a complexity measure, DIFFIT uses the sum of components. This difference results from the fact that the 'hull' heuristic was designed to select among T3, T2, T1, and CP solutions of different complexities, whereas DIFFIT considers T3 solutions only. Indeed, extending DIFFIT to CP, T2, and T1 solutions is not straightforward, since it is not clear how the *sum* concept can be generalized.
(2) Unlike the 'hull' heuristic, which was explicitly designed for selecting the set of solutions on the higher boundary of the convex hull of a plot of fit measure versus complexity measure, DIFFIT only approximates the set of 'hull' solutions.
(3) The implementation of the scree test in DIFFIT does not take into account the differences in the complexity of the solutions considered – that is, the differences in their *sum*-values.
(4) Unlike DIFFIT, the numerical convex hull based heuristic does not impose a minimum value for the *dif*-value of the selected solution.
(5) DIFFIT may select the simplest solution considered, whereas with the numerical convex hull based heuristic this is not possible.

6.2. Empirical comparison

Applying the numerical convex hull based heuristic to the 341 T3 solutions for the 225 randomly generated T3 data arrays of the simulation study (see Section 5.1) resulted in 12 selection errors, whereas applying DIFFIT resulted in 18 selection errors. Hence, it can be concluded that the more general 'hull' heuristic outperforms the DIFFIT method.

Given that we reported in Section 5.1 that the performance of the 'hull' heuristic is influenced by the size of a T3 data set, it could be conjectured that the advantage of the 'hull' heuristic over DIFFIT could be further increased by applying the 'hull' heuristic to sums of components rather than numbers of free parameters, as these sums do not depend on the size of a data set. To evaluate this conjecture, we applied the 'hull' heuristic to the *sum*-values (rather than the *fp*-values) and the *f*-values of the 341 T3 solutions for each of the 225 T3 simulated data sets mentioned above. It turned out that this alternative 'hull' heuristic worked even better, yielding only five selection errors. Unfortunately, this alternative 'hull' procedure cannot be used for selecting among different types of models, because generalizing the *sum*-concept to T2, T1, and CP models is not straightforward. However, given that this alternative 'hull' approach works so well for T3 data, it seems definitely recommendable for selecting among T3 models, and further research into its use for comparing models of different types seems indicated.

7. Discussion

In this paper, we have presented a numerical convex hull based model selection heuristic for selecting among three-mode principal component solutions of different types and complexities. Simulation results show that this heuristic performs almost perfectly, except for T3 data arrays with at least one small mode and a relatively large amount of error. Yet, it should be noted that a considerable number of three-way, three-mode arrays belong to this category. The reported simulation results, however, show that this problem can be tackled fairly well by applying the proposed procedure in a somewhat relaxed way, that is, also considering solutions close to the higher boundary of the convex hull or solutions with a high but not maximal *st*-value.

Our numerical convex hull based model selection heuristic may also be useful for solving other types of complex model selection problem. Indeed, promising results have already been reported for the family of three-mode hierarchical classes models (Ceulemans & Van Mechelen, 2005), a model family that is closely related to the family of principal component models. In particular, the proposed heuristic can be useful for all model selection problems for which a degrees-of-freedom-like measure and a fit measure are available for all solutions considered. Examples include loglinear analysis and structural equation modelling. However, as was also mentioned in Section 3.5 and holds for numerical model selection heuristics in general, one should not use this heuristic too rigidly, but rather as a helpful tool for making a first selection of interesting solutions. Indeed, for the final selection decision, one should also take into account substantive information and interpretability of the results.

Acknowledgements

The research reported in this paper was partially supported by the Research Council of K. U. Leuven (GOA/2000/02 and PDM/03/074). The first author is a post-doctoral fellow of the Fund for Scientific Research—Flanders (Belgium).

References

Bro, R., & Kiers, H. A. L. (2003). A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, *17*, 274–286.

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.

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

Fowlkes, E. B., Freeny, A. E., & Landwehr, J. M. (1988). Evaluating logistic models for large contingency tables. *Journal of the American Statistical Association*, *83*, 611–622.

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*, 1–84.

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

Kiers, H. A. L., & der Kinderen, A. (2003). A fast method for choosing the numbers of components in Tucker3 analysis. *British Journal of Mathematical and Statistical Psychology*, *56*, 119–125.

Kiers, H. A. L., & Harshman, R. A. (1997). Relating two proposed methods for speedup of algorithms for fitting two- and three-way principal component and related multilinear models. *Chemometrics and Intelligent Laboratory Systems*, *36*, 31–40.

Kroonenberg, P. M. (1983). *Three-mode principal component analysis: Theory and applications*. Leiden: DSWO.

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.

Kroonenberg, P. M., & Oort, F. J. (2003). Three-mode analysis of multimode covariance matrices. *British Journal of Mathematical and Statistical Psychology*, *56*, 305–336.

Kroonenberg, P. M., & Van der Voort, T. H. A. (1987). Multiplicatieve decompositie van interacties bij oordelen over de werkelijkheidswaarde van televisiefilms [Multiplicative decomposition of interactions for judgements of realism of television films]. *Kwantitatieve Methoden*, *8*, 117–144.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*, 661–675.

Murakami, T., & Kroonenberg, P. M. (2003). Three-mode models and individual differences in semantic differential data. *Multivariate Behavioral Research*, *38*, 247–283.

Timmerman, M. E., & Kiers, H. A. L. (2000). Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local minima. *British Journal of Mathematical and Statistical Psychology*, *53*, 1–16.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*, 279–311.

Verbeek, A. (1984). The geometry of model selection in regression. In T. K. Dijkstra (Ed.), *Misspecification analysis* (pp. 20–36). Berlin: Springer.

Wansbeek, T., & Verhees, J. (1989). Models for multidimensional matrices in econometrics and psychometrics. In R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 543–552). Amsterdam: North Holland.

Weesie, J., & Van Houwelingen, H. (1983). *GEPCAM users' manual (first draft)*. Utrecht, The Netherlands: Institute of Mathematical Statistics, State University of Utrecht.

Received 6 May 2004; revised version received 25 May 2005