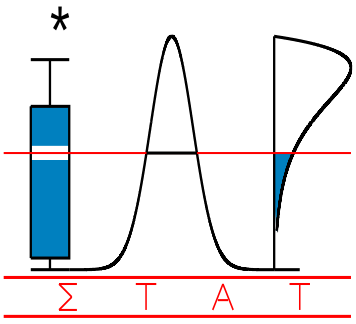
T E C H N I C A L

R E P O R T

0695

SELECTING AMONG MULTI-MODE PARTITIONING MODELS OF DIFFERENT COMPLEXITIES : A COMPARISON OF FOUR MODEL SELECTION CRITERIA

SCHEPERS, J., CEULEMANS. E. and I. VAN MECHELEN



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Selecting among multi-mode partitioning models of different complexities: a comparison of four model selection criteria

Jan Schepers, Eva Ceulemans and Iven Van Mechelen

Katholieke Universiteit Leuven

Author Notes:

J. Schepers and I. Van Mechelen were supported by the Fund for Scientific Research-Flanders (Belgium), Project No. G.0146.06 awarded to Iven van Mechelen and by the Research Council KULeuven (GOA/2005/04). E. Ceulemans is a post-doctoral fellow of the Fund for Scientific Research-Flanders (Belgium). Correspondence concerning this paper should be addressed to Jan Schepers, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium. Email: Jan.Schepers@psy.kuleuven.be

Running head: A comparison of four model selection criteria for multi-mode partitioning

Abstract

Multi-mode partitioning models for *N*-way *N*-mode data reduce each of the *N* modes in the data to a small number of clusters that are mutually exclusive. Given a specific *N*-mode data set, one may wonder which multi-mode partitioning model (i.e., with which numbers of clusters for each mode) yields the most useful description of this data set and should therefore be selected. In this paper, we address this issue by investigating four possible model selection heuristics: multi-mode extensions of Calinski and Harabasz's (1974) and Kaufman and Rousseeuw's (1990) indices for one-mode *k*-means clustering and multi-mode partitioning versions of Timmerman and Kiers's (2000) DIFFIT and Ceulemans and Kiers's (2006) numerical convex hull based model selection heuristic for three-mode principal component analysis. The performance of these four heuristics is systematically compared in a simulation study, which shows that the DIFFIT and numerical convex hull heuristics perform satisfactory in the two-mode partitioning case and almost perfectly in the three-mode partitioning case.

1. Introduction

Multi-mode partitioning models for *N*-way *N*-mode data reduce each of the *N* modes to a small number of clusters that are mutually exclusive. In particular, the two-mode partitioning model (Baier, Gaul & Schader, 1997; Gaul & Schader, 1996; Vichi, 2001) simultaneously partitions the rows and columns of a two-way two-mode data matrix, whereas the three-mode partitioning model (Kiers, 2004; Rocci & Vichi, 2005; Schepers, Van Mechelen & Ceulemans, 2006) simultaneously partitions the rows, columns, and slices of a three-way three-mode data array. As such, two-mode partitioning is an extension of the well-known *k*-means clustering method (e.g., Hartigan, 1975), which reduces one mode of a two-way two-mode data matrix to *k* non-overlapping clusters. Moreover, three-mode partitioning, apart from being a straightforward generalization of two-mode partitioning, is closely related to three-mode principal component analysis (3MCA; Tucker, 1966), which reduces each mode of a three-way three-mode data array to a small number of components. In particular, 3MCA reduces to three-mode partitioning if the component matrices for the three modes are constrained to take the form of partition matrices.

Up to now, most research in the domain of multi-mode partitioning has focused on the design and evaluation of algorithms for estimating multi-mode partitioning models (Castillo & Trejos, 2002; Schepers et al., 2006; van Rosmalen, Groenen, Trejos & Castillo, 2005).

A problem that has not yet received much attention is that of model selection: Given a two- or three-way data set, which multi-mode partitioning model (in terms of the number of clusters for each mode) yields the most useful description of the data set? In this paper we will address this model selection issue by discussing four possible model selection heuristics and by systematically comparing their performance in a simulation study. Given the close relation of multi-mode partitioning to traditional *k*-means clustering on the one hand and to

3MCA analysis on the other hand, two of the four considered model selection heuristics are multi-mode extensions of established criteria in traditional *k*-means clustering whereas the other two heuristics have originally been proposed to select among 3MCA models of different complexities.

The remainder of the paper is organized as follows: Section 2 recapitulates both two- and three-mode partitioning. Section 3 proposes the four model selection heuristics under consideration in the present paper in detail. In Section 4, a simulation study is presented in which the four considered heuristics are evaluated in terms of their capability of indicating the true complexity of two- and three-mode data sets. In Section 5 we applied each of these heuristics to set of mixed empirical-artificial data sets. Section 6 contains a few concluding remarks.

2. Multi-mode partitioning

2.1 Models

2.1.1 The two-mode partitioning model

The two-mode partitioning model approximates a real-valued $I \times J$ object by attribute data matrix **D** by a real-valued model or reconstructed data matrix **M** of the same size. Defining a partition matrix as a binary matrix of which each row sums to 1, **M** can be further decomposed into an $I \times P$ object partition matrix **A**, a $J \times Q$ attribute partition matrix **B** and a real-valued $P \times Q$ weight matrix **W**, with (P,Q) being the complexity of the model:

$$m_{ij} = \sum_{p=1}^{P} \sum_{q=1}^{Q} a_{ip} b_{jq} w_{pq}. \qquad (1)$$

In (1), a_{ip} and b_{jq} indicate whether or not object i and attribute j belong to object cluster p and attribute cluster q, respectively, and w_{pq} represents the strength of the relation between

clusters p and q. As such, (1) implies that m_{ij} equals w_{pq} iff object i and attribute j belong to object cluster p and attribute cluster q, respectively. In matrix notation, two-mode partitioning can be formalized as:

$$\mathbf{M} = \mathbf{AWB}'. \tag{2}$$

2.1.2 The three-mode partitioning model

Being a straightforward generalization of the two-mode partitioning model, the three-mode partitioning model approximates a real-valued $I \times J \times K$ object by attribute by source data array $\underline{\underline{\mathbf{D}}}$ by a real-valued model or reconstructed data array $\underline{\underline{\mathbf{M}}}$ of the same size, which can be further decomposed into $I \times P$, $J \times Q$, and $K \times R$ partition matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, and a real-valued $P \times Q \times R$ weight array $\underline{\underline{\mathbf{W}}}$, with (P,Q,R) being the complexity of the model:

$$m_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} w_{pqr}. \tag{3}$$

In matrix notation, the three-mode partitioning model is written as follows:

$$\mathbf{M}_\mathbf{A} = \mathbf{A}\mathbf{W}_\mathbf{A}(\mathbf{C}' \otimes \mathbf{B}'), \tag{4}$$

where $\otimes$ denotes the Kronecker product, and $\mathbf{M}_\mathrm{A}$ and $\mathbf{W}_\mathrm{A}$ denote the $I \times JK$ and $P \times QR$ matricizations of $\underline{\underline{\mathbf{M}}}$ and $\underline{\underline{\mathbf{W}}}$ respectively (Kiers, 2000).

2.2 Data analysis

Given a multi-mode data array $\underline{\underline{\mathbf{D}}}$ and a desired complexity (i.e., a prespecified number of clusters for each mode of $\underline{\underline{\mathbf{D}}}$), multi-mode partitioning looks for a model array $\underline{\underline{\mathbf{M}}}$ of the desired complexity that maximizes the percentage of data variance accounted for by the model:

$$VAF = \left(1 - \frac{\| \underline{\underline{\mathbf{D}}} - \underline{\underline{\mathbf{M}}} \|^2}{\| \underline{\underline{\mathbf{D}}} - \underline{\underline{\bar{d}}} \|^2}\right) \times 100, \tag{6}$$

where $\bar{d}$ and ||.|| denote the overall data mean and the Euclidean norm, respectively. To obtain multi-mode partitioning solutions with good *VAF*-values, several algorithms have been proposed. In particular, Gaul and Schader (1996), Baier et al. (1997), Castillo and Trejos (2002), and van Rosmalen et al. (2005) proposed algorithms for the two-mode case, whereas Rocci and Vichi (2005), Kiers (2004) and Schepers et al. (2006) proposed three-mode partitioning algorithms. For the analyses reported in this paper, we used the three-mode partitioning algorithm (i.e., DRIFT with 50 random multi-starts) that turned out to be most robust in the comparative simulation study performed by Schepers et al. (2006). Note that any three-mode partitioning algorithm can also be used for two-mode partitioning of data matrices, because a two-mode data matrix can also be conceived as a three-mode data array with the third mode consisting of one element only.

3. Four model selection heuristics for multi-mode partitioning

Up to now, almost no heuristics for selecting the numbers of clusters in multi-mode partitioning have been proposed in the literature. One exception to this is an extension of the well-known Calinski-Harabasz index (Rocci and Vichi, 2005) (see below). Promising candidate heuristics could be inspired by model selection work in the domain of one-mode partitioning on the one hand and in the domain of multi-mode component analysis on the other hand. In this section we will discuss four promising model selection heuristics: Two are derived from existing heuristics for one-mode partitioning and two from methods for selecting among multi-mode component analysis models of different complexities.

3.1 Extended Calinski-Harabasz indices

In order to select the most useful number of clusters in a traditional *k*-means clustering of the objects of a given object by attribute data matrix, Calinski and Harabasz (1974) proposed the following *CH*–index:

$$CH = \frac{trace(\mathbf{B})/(k-1)}{trace(\mathbf{W})/(I-k)}. \quad (7)$$

As *trace*(**W**) and *trace*(**B**) are the within- and between-cluster sums-of-squares, and *k* and *I* denote the number of clusters and the total number of objects, it can be concluded that the *CH*-index divides the between-cluster variability of a solution by the corresponding within-cluster variability, both corrected for their respective degrees of freedom. As such, the *CH*-index will attain high values for solutions with enough but not too many clusters. The Calinski-Harabasz model selection heuristic then consists of two steps:

1. For all *k*-means clustering solutions among which one wants to select, determine the value of the *CH*-index. Note, however, that *CH* is not defined for *k* = 1, implying that one will never select a model with one cluster only.
2. Select the solution with the highest CH-value.

The Calinski-Harabasz index outperformed 29 other *k*-means clustering model selection heuristics in Milligan and Cooper's (1985) well-known simulation study. Therefore, it is not surprising that this index still is the most widely used and preferred model selection criterion in traditional *k*-means clustering.

Rocci and Vichi (2005) already proposed a two-mode partitioning extension of the Calinski-Harabasz index; however, no systematic investigation of its performance has been pursued so far. In particular, Rocci and Vichi (2005) proposed to choose the solution for which it holds that

$$ExtCH_2 = \frac{\| \mathbf{AWB}' - \bar{d} \|^2 / (PQ-1)}{\| \mathbf{D} - \mathbf{AWB}' \|^2 / (IJ-PQ)}, \quad (8)$$

is maximal. It can be derived that, like *CH*, $ExtCH_2$ divides the sum-of-squares explained by the model by the residual sums-of-squares, both corrected for their respective degrees of freedom. The extended two-mode Calinski-Harabasz index can easily be generalized to the three-way case:

$$ExtCH_3 = \frac{\| \mathbf{AW_A}(\mathbf{C}'\otimes\mathbf{B}') - \bar{d} \|^2 /(PQR-1)}{\| \mathbf{D_A} - \mathbf{AW_A}(\mathbf{C}'\otimes\mathbf{B}') \|^2 /(IJK-PQR)}, \qquad (9)$$

It is easily verified that in case $K = R = 1$, (9) is equivalent to (8). Note that $ExtCH_2$ and $ExtCH_3$ are undefined for $P = Q = 1$ and $P = Q = R = 1$, respectively, implying that one will never select a multi-mode partitioning solution with one cluster only for each mode.

3.2 Extended silhouette indices

Kaufman and Rousseeuw (1990) proposed the Silhouette index for choosing among *k*-means clustering solutions with varying numbers of clusters. This index is defined for a given object i as

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \qquad (10)$$

where a_i and b_i denote the average (Euclidean) distance of object *i* to the other objects in the same cluster, and the average (Euclidean) distance of object *i* to the objects in the nearest other cluster, respectively, where the nearest other cluster is defined as the one with the minimal b_i-value. Clearly, the numerator of (10) becomes large, if the solution contains at least enough clusters to adequately capture the variability in the objects. By dividing this numerator by the maximum of a_i and b_i, the index necessarily takes values between -1 and 1, indicating poor and well clustering, respectively, for object *i.* Note that the concept of other clusters only makes sense if $k>1$, implying that solutions with one cluster only are not considered. The average of the Silhouette-index S_i across objects will further act as a model selection heuristic, with the solution with the highest average Silhouette-value being selected.

Recent studies (Tibshirani, Walther & Hastie, 2001; Sugar & James, 2003) showed that the Silhouette index performs almost as good as the Calinski-Harabasz index.

In this paper, we will evaluate the following two-mode extension of the Silhouette heuristic: For all two-mode partitioning solutions under consideration, we determine the value of the Silhouette-index for each object and each attribute separately. Note that these indices can only be computed if $P > 1$ and $Q > 1$; in other words, only solutions with more than one cluster for each mode are considered. Next we compute the average Silhouette-index across objects $E(S_1)$ and the average Silhouette index across attributes $E(S_2)$, which are in their turn averaged, weighted by the number of objects and attributes:

$$ExtS_2 = \frac{I \times E(S_1) + J \times E(S_2)}{I + J} \tag{11}$$

The solution that maximizes $ExtS_2$ is to be selected.

For selecting among three-mode partitioning models, the procedure is similar, making use of a weighted combination ($ExtS_3$) of the average silhouette indices for the objects, attributes and sources.

3.3 DIFFIT

Both this method and the next one are based on the idea that the most promising 3MCA solutions, that is, the solutions with the best fit-complexity balance, are located at elbows in the upper boundary of the convex hull of a goodness-of-fit versus complexity plot of the set of solutions from which one wants to choose (Kroonenberg & Van der Voort, 1987; Kroonenberg & Oort 2003). This convex hull idea implies that the complexity of 3MCA solutions is to be summarized by means of a single number instead of three (i.e., the number of components for each mode). In particular, the DIFFIT method (Timmerman & Kiers, 2000) works as follows:

1. For all 3MCA solutions among which one wants to select, determine the *VAF*-values. Furthermore, the complexity of the different solutions is expressed by one number *sum*, that is, the sum of the numbers of components $P+Q+R$.
2. For each observed *sum*-value, retain only the best fitting solution, that is, the solution with the highest *VAF*-value. Indicate the N retained solutions by s_{sum}.
3. For each of the N retained solutions, compute dif_{sum} as the difference between VAF_{sum} and VAF_{sum-1}; this implies that the *dif*-value of the simplest solution (*1,1,1*) equals 0.
4. Exclude all solutions s_i for which a solution s_j ($j>i$) exists such that $dif_j>dif_i$. Number the M remaining solutions by m=1..M. The associated *sum*-values are given by *sum(m)*, implying that the corresponding *dif*-values are given by $dif_{sum(m)}$.
5. For the first *M-1* solutions, compute

$$b_{sum(m)}=dif_{sum(m)}/dif_{sum(m+1)}. \quad (12)$$

6. Eliminate solutions that entail *VAF* increases that are smaller than the overall expected increase in *VAF* when considering a more complex solution, the latter being estimated by $\|\underline{\mathbf{X}}\|^2/(sum_{max}-3)$, with sum_{max}=min(*I,JK) +min(J,IK) + min(K,IJ)*. Note that the denominator, sum_{max} - 3, is based on the observation of Wansbeek and Verhees (1989) that sum_{max} is the highest sensible complexity and that three *sum* values (i.e., 1, 2, and 4) will not occur or are not sensible.
7. From the remaining solutions, select the solution with the highest $b_{sum(m)}$-value.

The DIFFIT heuristic showed almost perfect performance when evaluated in simulation studies. To obtain two- and three-mode partitioning versions of DIFFIT, the following simple adjustments suffice: In step 1, to express the complexity of a two-mode partitioning solution by a single number, the sum $P+Q$ of the number of clusters for the two

modes is computed; similarly, the complexity of a three-mode partitioning solution is given by $P+Q+R$. With respect to step 6, for multi-mode partitioning, sum_{max} equals the sum of the numbers of elements pertaining to each mode. Furthermore, whereas in case of two-mode partitioning, only a *sum*-value of 1 can not occur, in three-mode partitioning *sum*-values of 1 and 2 are not sensible; therefore, we divide $\|\underline{\mathbf{X}}\|^2$ by $I+J-1$ and $I+J+K-2$, respectively. One may note that, in contrast to the extended Calinski-Harabasz and Silhouette indices, DIFFIT allows to make a selection among all possible solutions.

3.4 Numerical convex hull method

Ceulemans and Kiers (2006) presented the following numerical convex hull based method, to select among 3MCA solutions of different complexities:

1. Determine the complexity- and *VAF*-values of all 3MCA solutions from which one wants to choose. Ceulemans and Kiers (2006) considered two options for quantifying the complexity of a solution: the number of free parameters $fp=IP+JQ+KR+PQR-P^2-Q^2-R^2$ and the sum of components $sum=P+Q+R$. Simulation results showed that, for choosing among 3MCA solutions only, the use of the *sum*-values yielded slightly better results than the use of the *fp*-values.
2. For each of the n observed *sum*-values, retain only the best fitting solution, that is, the solution with the highest *VAF*-value.
3. Sort the n retained solutions by their *sum*-values and denote them by s_i (i=1…n).
4. Exclude all solutions s_i for which a solution s_j ($j<i$) exists such that $VAF_j > VAF_i$.
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 *VAF* versus *sum* plot.

6. Repeat Step 5 until no solution can be excluded anymore. This step yields the solutions on the upper boundary of the convex hull.
7. Determine the *st*-values

$$st_i = \frac{VAF_i - VAF_{i-1}}{sum_i - sum_{i-1}} \Big/ \frac{VAF_{i+1} - VAF_i}{sum_{i+1} - sum_i}. \quad (13)$$

   of the retained convex hull solutions.
8. Select the solution with the highest *st*-value. Note that a relatively large *st*-value indicates that allowing for sum_i components (instead of sum_{i-1} components) increases the *VAF*-value of the model considerably, whereas allowing for more than sum_i components hardly increases the *VAF*-value. Thus, the solution is selected after which the increase in *VAF*-value levels off. Note that it is impossible to select the least and the most complex model, respectively, since (13) is undefined for these solutions.

This method appeared to have almost perfect performance in simulation studies. Moreover, it also performed well when used for selecting among three-way hierarchical classes models of different complexities (Ceulemans & Van Mechelen, 2005), the models in question implying an overlapping clustering of all modes in a three-way three-mode binary data array.

To obtain multi-mode partitioning versions of the numerical convex hull procedure, it is only necessary to specify how to quantify the complexity of a solution in Step 1. In this regard, we propose to use for a two-mode partitioning solution the sum $P+Q$ of the numbers of clusters for the two modes, and for a three-mode partitioning the sum $P+Q+R$ of the numbers of clusters for each of the three modes.

4. Simulation study

In this section, we present two simulation studies in which we evaluate to which extent the different model selection criteria succeed in indicating the true underlying complexity in the data. First, we discuss the design of the simulation study and the performance of the four model selection criteria with respect to indicating the correct underlying numbers of clusters in the context of choosing among two-mode partitioning models. Next, we discuss the design and the results pertaining to a simulation study in which the selection of the numbers of clusters of three-mode partitioning models for three-way three-mode data is investigated.

4.1 Two-mode partitioning

4.1.1 Design

To explain the design of the data generation, three different types of real-valued $I \times J$ matrices must be distinguished: a true matrix **T**, which can be represented by a two-mode partitioning model with a prespecified number of clusters for each mode; a data matrix **D**, which is **T** perturbed with error; and the model matrix **M** yielded by the two-mode partitioning estimation, which can be represented by a two-mode partitioning model with the same number of underlying clusters as the true matrix **T**. Three design factors were fully crossed on the level of the data generation (matrices **T** and **D**) for the data generated in this simulation study:

1. the size $I \times J$ of the data matrices at two levels: 40 × 40, 80 × 20;
2. the numbers of clusters present in the two-mode partitioning model that underlies the data matrices, at five levels: (2,2), (3,2), (3,3), (4,3), (4,4);
3. the amount of error on the data at three levels: 15, 30, 45%.

For each cell of the design ten replications were considered, yielding 300 two-mode simulated data sets. In particular, for each combination of the levels of size, numbers of clusters and error, partition matrices **A** and **B** are drawn by randomly assigning each element of a mode to

one of its corresponding clusters. A true matrix **T** then is obtained by generating entries of **W** as independent realizations of a uniformly distributed variable in [0,1] and by combining the resulting **A**, **B** and **W** by (1). Next, corresponding data matrices **D** are obtained by

$$\mathbf{D} = \mathbf{AWB}' + \varepsilon\mathbf{E} = \mathbf{T} + \varepsilon\mathbf{E}, \qquad (14)$$

where 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 $\|\mathbf{T}\| = \|\mathbf{E}\|$.

Each of these 300 simulated data sets were subjected to 25 two-mode partitionings by considering all combinations of numbers of clusters ranging from (1,1) to (5,5). Subsequently, for each data set, each of the four model selection criteria were applied to the obtained 25 solutions.

4.1.2 Results

Figure 1 displays the frequencies with which each of the four model selection criteria indicate the correct underlying numbers of clusters. The numerical convex hull procedure performs best with 236 ‘hits’ out of 300 (or 79%). The DIFFIT method performs almost as good with 234 hits (78%). The Extended Calinski-Harabasz and Silhouette indices clearly do not perform as well as the former two methods (with hit percentages of 25 and 30, respectively).

[insert Figure 1 about here]

For 57 out of the total of 300 data sets (19%), both DIFFIT and the numerical convex hull select the same incorrect solution. A closer examination of these data sets reveals an underestimation of the underlying numbers of clusters in each case, that is, the indicated number of clusters for one mode is never larger than the true number of clusters for that

mode. Moreover, it appeared that the corresponding *VAF*-values of these incorrectly chosen solutions in general are almost as large as the goodness-of-data values (*GOD*) which indicate the percentage of variance accounted for in the data by the true underlying model:

$$GOD = \left(1 - \frac{\| \mathbf{D} - \mathbf{T} \|^2}{\| \mathbf{D} - \bar{d} \|^2}\right) \times 100, \qquad (15)$$

This means that the "incorrectly" chosen solutions explain a fairly large amount of the structural part in the data and in this respect act as approximately equivalent to the true underlying model.

In order to investigate the effect of the three design factors (size, numbers of clusters and error) on the performance of each of the four model selection methods, we counted how often each method in each condition indicated the correct underlying numbers of clusters. We analyzed these frequencies (between 0 and 10) by means of a repeated multivariate measures analysis of variance (RMANOVA), using only main effects and first-order interactions of model selection method and the independent variables. Sufficiently large effect sizes ($\eta^2 > 0.10$) are only observed for numbers of clusters ($\eta^2 = 0.20$), model selection method ($\eta^2 = 0.51$) and interaction between numbers of clusters and model selection method ($\eta^2 = 0.11$). It can be seen from Figure 2 that the sizeable interaction is mostly due to a large decrease in performance (i.e., from perfect to worst out of four) for the Extended Silhouette index when the numbers of clusters becomes larger than (2,2). Two remarks are important here: First, as discussed in 3.2, the Extended Silhouette index does not allow the selection of models with less than two clusters for each mode. This implies that in all cases where the true underlying numbers of clusters equals (2,2), a correct selection corresponds to the selection of the least complex model possible for this index. Second, closer inspection of the cases where this index selects an incorrect solution shows that it always underestimates the underlying numbers of clusters. Taking into account these two remarks, it is not surprising that this index performs

well only when the true numbers of clusters equals (2,2). Finally, we note that the same tendency to underestimate the numbers of clusters is observed for the Extended Calinski-Harabasz index.

[insert Figure 2 about here]

4.2 Three-mode partitioning

4.2.1 Design

For the data generated in the part of the simulation study that pertains to three-mode partitioning, a completely analogous design was used as for two-mode partitioning. The levels of the three design factors here are:

1. the size $I \times J \times K$ of the data arrays at three levels: 20 × 20 × 20, 30 × 30 × 9, 80 × 10 × 10;
2. the true numbers of clusters present in the three-mode partitioning model that underlies the data arrays, at five levels: (2,2,2), (3,2,2), (4,2,2), (4,3,2), (4,4,4);
3. the amount of error on the data at three levels: 15, 30, 45%.

By generating 10 replicates in each cell of the design by $\mathbf{D}_{\mathbf{A}} = \mathbf{A}\mathbf{W}_{\mathbf{A}}(\mathbf{C}' \otimes \mathbf{B}') + \varepsilon\mathbf{E}$, a total of 450 data sets are considered in this simulation study.

Each of these 450 simulated data sets were subjected to 125 three-mode partitioning analyses by considering all combinations of numbers of clusters ranging from (1,1,1) to (5,5,5). Subsequently, for each data set, each of the four model selection criteria was applied to the resulting 125 solutions.

4.2.2 Results

As depicted in Figure 3, for 124 (28%), 196 (44%), 438 (97%) and 442 (98%) out of the total of 450 data sets, the Extended Calinski-Harabasz index, the Extended Silhouette index, the DIFFIT procedure and the numerical convex hull heuristic, respectively, succeeded in indicating the correct underlying numbers of clusters.

[insert Figure 3 about here]

Clearly, as in the two-mode case, DIFFIT and the numerical convex hull heuristic are superior to the Extended Calinski-Harabasz and Silhouette indices. Moreover, the former two model selection heuristics perform almost perfect. We used the same repeated multivariate measures analysis of variance procedure as in Section 4.1.2, using only main effects and first-order interactions of model selection method and the independent variables. A sufficiently large effect size ($\eta^2 > 0.10$) is only observed for model selection method ($\eta^2 = 0.68$), indicating that the design factors are more or less irrevelant in explaining the performance in indicating the correct numbers of clusters. As an aside, it is interesting to note that, similarly to the two-mode case, the Extended Calinski-Harabasz index, whenever it selected an incorrect three-mode partitioning solution, always underestimated the correct numbers of clusters. No such systematic (mis)behavior was observed for the Extended Silhouette index.

5. Application to a collection of mixed empirical-artificial data sets

One may wonder to which extent the results found in the simulation studies above also hold for more realistic data sets. In particular, the question may be raised whether our reported findings apply in cases where (1) the underlying structure is a natural one (as opposed to an

artificially simulated one), and (2) the error on the data is not Gaussian distributed. In this section, we will apply all four model selection heuristics to a collection of simulated data sets, all of which are constructed on the basis of an empirical data set, 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 twenty bipolar scales (e.g., bright-dark, slow-fast). Murakami and Kroonenberg (2003) suggested to preprocess the resulting 24 (preludes) × 20 (scales) × 38 (participants) data array **D** by centring the scores across the prelude mode and by subsequently normalizing the scores per scale.

We analyzed this data set with the numbers of clusters ranging from (1,1,1) to (5,5,5), yielding 125 three-mode partitioning solutions. Subsequently, on the basis of the numerical convex hull method we retained the solution with numbers of clusters equal to (2,3,2) (the DIFFIT method indicates the same solution as most appropriate).

Next, we obtained 100 simulated data sets from the Chopin data and the solution with complexity (2,3,2) by: (a) reconstructing the model array **M** according to (3), (b) obtaining the residual array **R** by subtracting **M** from **D**, and (c) constructing 100 simulated data arrays by adding residual arrays, the entries of which were drawn randomly with replacement from the set of residuals **R**, to the model array **M**. Note that the 100 simulated three-way three-mode data arrays constructed in this way have the same empirically obtained underlying true structure, and do not include iid Gaussian distributed error.

We analyzed all 100 mixed empirical-artificial data sets, with the numbers of clusters ranging from (1,1,1) to (5,5,5), yielding 125 three-mode partitioning solutions for each. Next, we applied each of the four model selection heuristics considered in this paper to the set of solutions for each data set and calculated the proportion of data sets for which each of the heuristics indicated the "true" underlying numbers of clusters of (2,3,2). These proportions

equalled 0, .81, 1.00 and 1.00 for the Extended Calinski-Harabasz index, the Extended Silhouette index, the DIFFIT procedure and the numerical convex hull heuristic, respectively. In line with the findings of our two simulation studies with artificial data, the DIFFIT and convex hull methods have a perfect performance. The performance of the extended Calinski-Harabasz index is even worse than in the simulation studies with artificial data whereas the extended Silhouette index seems to perform somewhat better than expected. We conclude that the ordering of the four model selection heuristics on the basis of their model selection performance for this set of 100 mixed empirical-artificial data sets is the same as the one that showed up in our simulation studies with artificial data.

6. Concluding remarks

In this paper we compared the performance of four different model selection heuristics to select among multi-mode partitioning models of different complexities. It was found that the procedures specifically designed for a multi-mode context (i.e., DIFFIT and numerical convex hull) perform much better than procedures based on an extension of what usually is applied in one-mode *k*-means clustering (i.e., Calinski-Harabasz and Silhouette). Moreover, it was found that the performance of the former two model selection criteria ranged from satisfactory (in the two-mode case) to almost perfect (in the three-mode case). The difference between the model selection methods further was found to be the most important factor in explaining the performance in indicating the correct numbers of clusters. Only in the two-mode case, the (interaction with) numbers of clusters also turned out to be important. However, this finding could be attributed to the fact that the Extended Silhouette index does not allow to select models with less than two clusters for each mode.

The multi-way model selection criteria differ from the extensions of one-mode procedures in two respects: First, they explicitly take into account the complexity of the considered solutions. Second, the logic behind the numerical convex hull, and implicitly also the DIFFIT method, is that only those solutions are considered that yield an increase in explained variance that is not smaller than increases for subsequent solutions (i.e., the solutions that lie on the upper boundary of the convex hull). Note further that multi-way model selection criteria require a specification of the model complexity. Therefore, both aspects cannot easily be disentangled in an explanation of the good performance of the multi-way model selection criteria.

The other two methods considered in this paper are merely straightforward extensions of model selection criteria that have been shown to be useful in the context of one-mode *k*-means clustering. As such, one may argue that these methods, unlike the DIFFIT and numerical convex hull methods, do not fully take into account the specific nature of multi-mode data and models.

References

Baier, D., Gaul, W., & Schader, M. (1997). Two-mode overlapping clustering with applications to simultaneous benefit segmentation and market structuring. In: R. Klar, & O. Opitz (Eds.), *Classification and knowledge organization*. Springer, Berlin, 557-566.

Calinski, R.B., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*, 1-27.

Castillo, W., & Trejos, J. (2002). Two-mode partitioning: Review of methods and application of tabu search. In: K. Jajuga, A. Sokolowski, & H.-H. Bock (Eds.), *Classification, clustering and related topics: Recent advances and applications*. Springer, Heidelberg, 43-51.

Ceulemans, E., & Kiers, H.A.L. (2006). Selecting among three-way principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology, 59,* 133-150.

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

Gaul, W., & Schader, M. (1996). A new algorithm for two-mode clustering. In: H.-H. Bock, & W. Polasek (Eds.), Classification *and knowledge organization*. Springer, Berlin, 15-23.

Hartigan, J. (1975). *Clustering algorithms*. New York: Wiley.

Kauffman, L., & Rousseeuw, P. (1990). *Finding groups in Data: an introduction to cluster analysis*. New York: Wiley

Kiers, H.A.L. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics, 14*, 105-122.

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

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.

Milligan, G.W., & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159-179.

Rocci, R. & Vichi, M. (2005). Three-mode component analysis with crisp or fuzzy partition of units. *Psychometrika, 70*, 715-736.

Schepers, J., Van Mechelen, I. & Ceulemans, E. (2006). Three-mode partitioning. *Computational Statistics & Data Analysis*, doi:10.1016/j.csda.2006.06.002

Sugar, C.A., & James, G.M. (2003). Identifying groups in data: An information-theoretic approach. *Journal of the American Statistical Association, 98*, 750-763.

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.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B, 63*, 411-423.

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

van Rosmalen, J., Groenen, P.J.F., Trejos, J., & Castillo, W. (2005). Global optimization strategies for two-mode clustering. Econometric Institute Report EI 2005-33.

Vichi, M. (2001). Double *k*-means clustering for simultaneous classification of objects and variables. In: S. Borra, R. Rocci, & M. Schader (Eds.), *Advances in classification and data analysis*, Springer, Heidelberg, 43-52.

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.

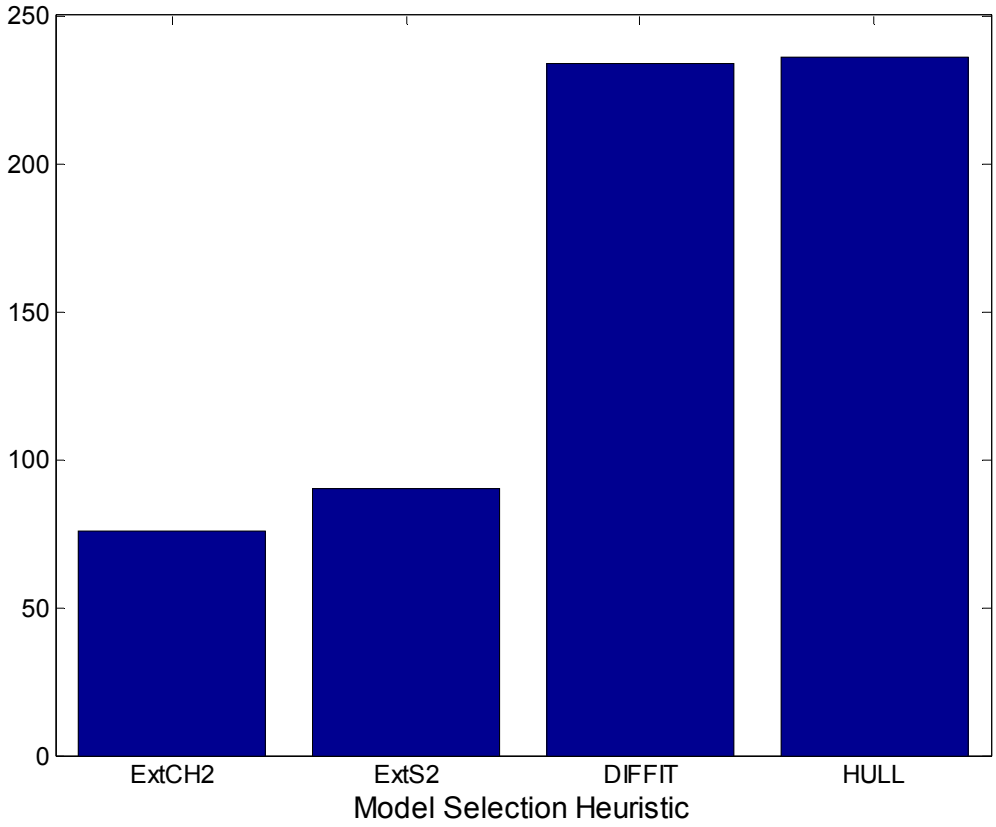

250
200
150
100
50
0
ExtCH2
ExtS2
DIFFIT
HULL
Model Selection Heuristic

Figure 1.

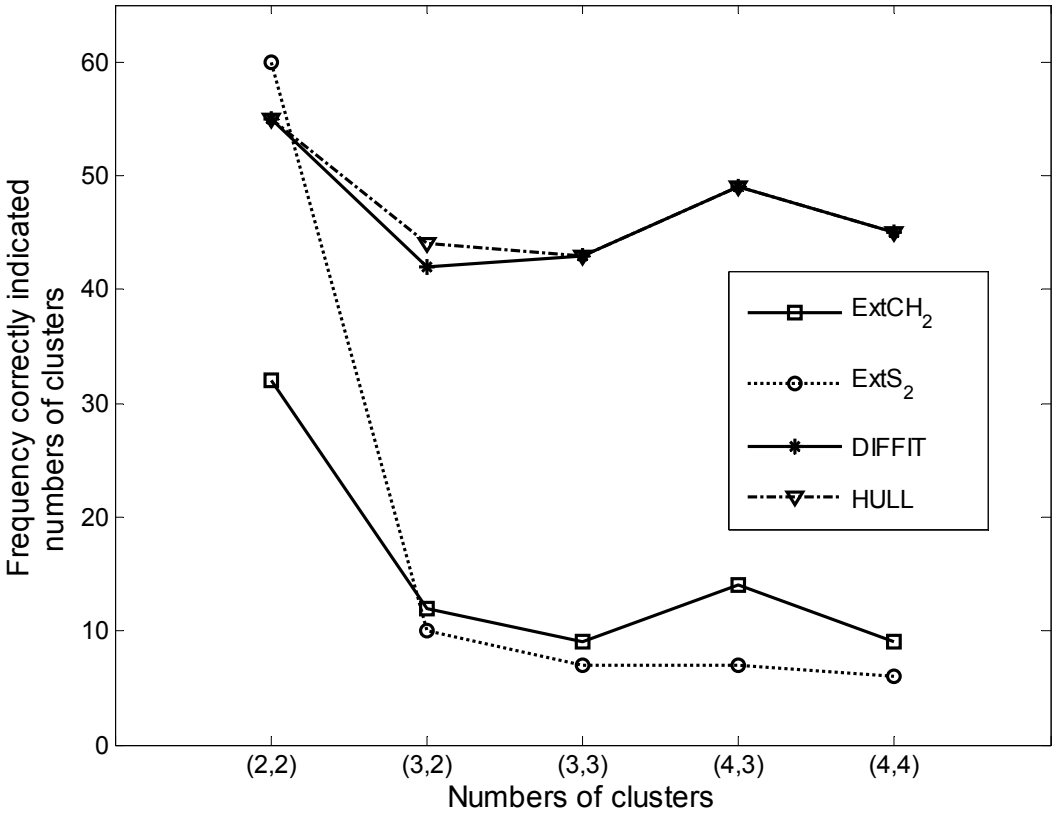

Frequency correctly indicated numbers of clusters
60
50
40
30
20
10
0
(2,2)
(3,2)
(3,3)
(4,3)
(4,4)
Numbers of clusters
ExtCH2
ExtS2
DIFFIT
HULL

Figure 2.

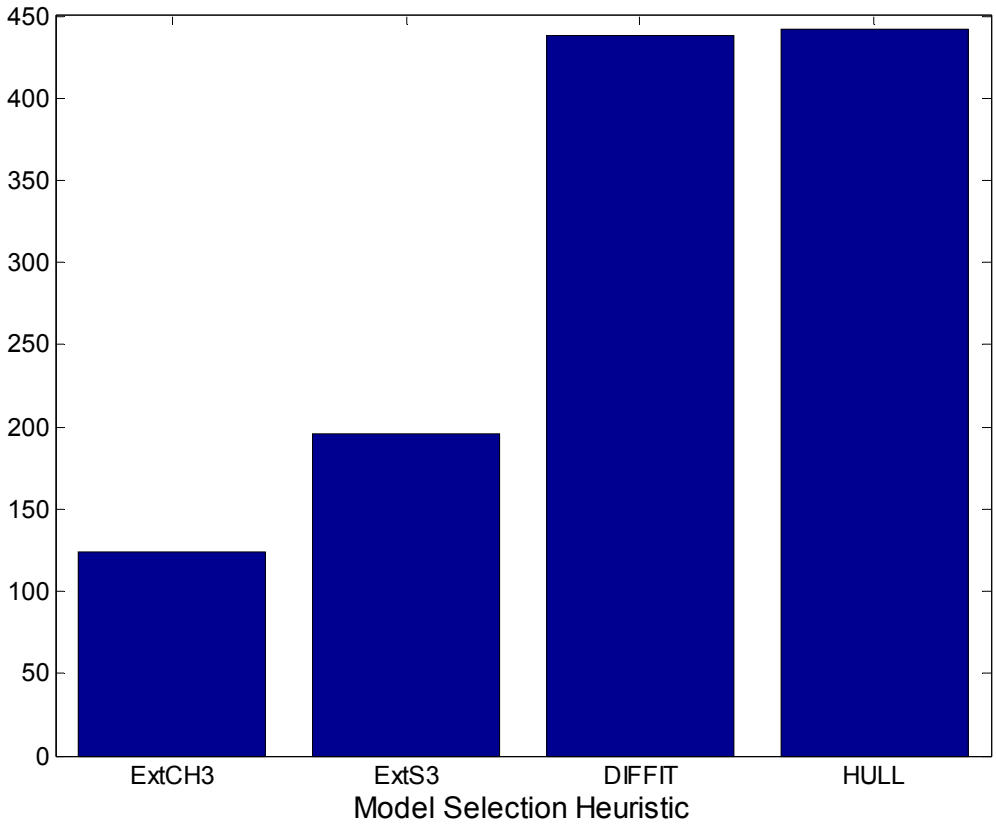

450
400
350
300
250
200
150
100
50
0
ExtCH3
ExtS3
DIFFIT
HULL
Model Selection Heuristic

Figure 3.

Figure captions.

Figure 1. Frequency of correct two-mode model selection by each of the four model selection criteria under study.

Figure 2. Mean frequency of correct two-mode model selection by each of the four model selection criteria under study, for each level of numbers of clusters.

Figure 3. Frequency of correct three-mode model selection by each of the four model selection criteria under study