Journal of Classification 25:67-85 (2008)
DOI: 10.1007/s00357-008-90005-9

Selecting Among Multi-Mode Partitioning Models of Different Complexities: A Comparison of Four Model Selection Criteria

Jan Schepers

Katholieke Universiteit Leuven, Belgium

Eva Ceulemans

Katholieke Universiteit Leuven, Belgium

Iven Van Mechelen

Katholieke Universiteit Leuven, Belgium

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 very good in the three-mode partitioning case.

Keywords: Multi-mode partitioning; Model selection; Number of clusters.

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).

Corresponding Author's Address: Jan Schepers, Katholieke Universiteit Leuven, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium, e-mail: Jan.Schepers@psy.kuleuven.be

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 and Schadder 1997; Gaul and Schader 1996; Vichi 2002) partitions the rows and columns of a two-way two-mode data matrix, whereas the three-mode partitioning model (Kiers 2004; Schepers, Van Mechelen and 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 (MacQueen 1967; 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 and Trejos 2002; Schepers et al. 2006; van Rosmalen, Groenen, Trejos and 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 a 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 $\mathbf{D}$ by a real-valued model or reconstructed data matrix $\mathbf{M}$ of the same size. Defining a partition matrix as a binary matrix of which each row sums to 1, $\mathbf{M}$ can be decomposed into an $I \times P$ object partition matrix $\mathbf{A}$, a $J \times Q$ attribute partition matrix $\mathbf{B}$ and a real-valued $P \times Q$ weight matrix $\mathbf{W}$, with (P, Q) being the complexity of the model:

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

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 $\underline{\mathbf{D}}$ by a real-valued model or reconstructed data array $\underline{\mathbf{M}}$ of the same size, which can be 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 $\underline{\mathbf{W}}$, 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:

$$\mathbf{M}_A = \mathbf{AW}_A(\mathbf{C}' \otimes \mathbf{B}'), \tag{4}$$

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

2.2 Data Analysis

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

$$VAF = \left(1 - \frac{||\underline{\mathbf{D}} - \underline{\mathbf{M}}||^2}{||\underline{\mathbf{D}} - \bar{d}||^2}\right) \times 100, \tag{5}$$

where $\bar{d}$ and $||.||$ denote the overall data mean and the Euclidean norm, respectively. Several algorithms have been proposed to tackle this optimization problem. In particular, Gaul and Schader (1996), Baier, Gaul, and Schader (1997), Castillo and Trejos (2002), and van Rosmalen, Groenen, Trejos, and Castillo (2005) proposed algorithms for the two-mode case, whereas Rocci and Vichi (2008), Kiers (2004) and Schepers, Van Mechelen, and Ceulemans (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). In particular, this algorithm starts from a random assignment of each element for each mode to one of the clusters for that mode; and subsequently the cluster assignments for each mode are updated in an alternating way, until the sum of squared differences between $\underline{\mathbf{D}}$ and $\underline{\mathbf{M}}$ no longer decreases. For the updating of cluster assignments within one mode the algorithm starts with the first element of that mode, evaluates whether reassigning it to a different cluster decreases the discrepancies between $\underline{\mathbf{D}}$ and $\underline{\mathbf{M}}$ (taking into account corresponding changes in $\mathbf{W}$), and subsequently updates the cluster assignment of the following elements. This procedure is then repeated until no further changes in cluster assignment have occurred within that mode. (For further details we refer to 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 2008) (see below). Promising candidate heuristics could be inspired by model selection work in the domain of one-mode partition-

ing 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 that makes use of the within-group dispersion matrix **WG** (**WG** $= 1/(n-k)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_j)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_j)'$) and between-group dispersion matrix **BG** (**BG** $= \sum_{i=1}^{k} n_i(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})'$), where $\mathbf{x}_{ij}$ is the vector of attribute values for the j-th object in the i-th group, $\bar{\mathbf{x}}$ is the mean vector of all n objects, $\bar{\mathbf{x}}_i$ is the mean vector of the objects in group i, and n_i is the number of objects in group i (Everitt 1993):

$$CH = \frac{trace(\mathbf{BG})/(k-1)}{trace(\mathbf{WG})/(n-k)}. \tag{6}$$

As trace(**WG**) and trace(**BG**) are the within- and between-cluster sums-of-squares, and k and n 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 indices for determining the number of clusters in a data set in Milligan and Cooper's (1985) well-known simulation study on stopping rules for hierarchical clustering. Therefore, it is not surprising that this index still is the most widely used and preferred model selection heuristic in the clustering domain. Rocci and Vichi (2008) have 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 (2008) proposed to choose the solution for which it holds that

$$ExtCH_2 = \frac{||\mathbf{AWB'} - \bar{d}||^2/(PQ-1)}{||\mathbf{D} - \mathbf{AWB'}||^2/(IJ-PQ)} \tag{7}$$

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:

$$ExtCH_3 = \frac{||\mathbf{AW}(\mathbf{C'} \otimes \mathbf{B'}) - \bar{d}||^2/(PQR-1)}{||\mathbf{D}_A - \mathbf{AW}_A(\mathbf{C'} \otimes \mathbf{B})||^2/(IJK-PQR)}. \tag{8}$$

It is easily verified that in case $K = R = 1$, (8) is equivalent to (7). 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

Kauffman 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

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{9}$$

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 (9) 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 good 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 and Hastie 2001; Sugar and 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:

$$ExtS_2 = \frac{I \times E(S_1) + J \times E(S_2)}{I + J}. \tag{10}$$

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:

$$ExtS_3 = \frac{I \times E(S_1) + J \times E(S_2) + K \times E(S_3)}{I + J + K}. \tag{11}$$

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 and Van der Voort 1987; Kroonenberg and 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 and 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

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

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 $||X||^2/(sum_{max}-3)$, 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 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 also allows to select solutions with one component for any or even all modes.

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
$$st_i = \frac{VAF_i - VAF_{i-1}}{sum_i - sum_{i-1}} \Big/ \frac{VAF_{i+1} - VAF_i}{sum_{i+1} - sum_i} \tag{13}$$
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 and 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. Figure 1 displays the major steps in the numerical convex hull heuristic for a three-mode data set.

4. Simulation Study

In this section, we present two simulation studies in which we evaluate to which extent the different model selection heuristics succeed in in-

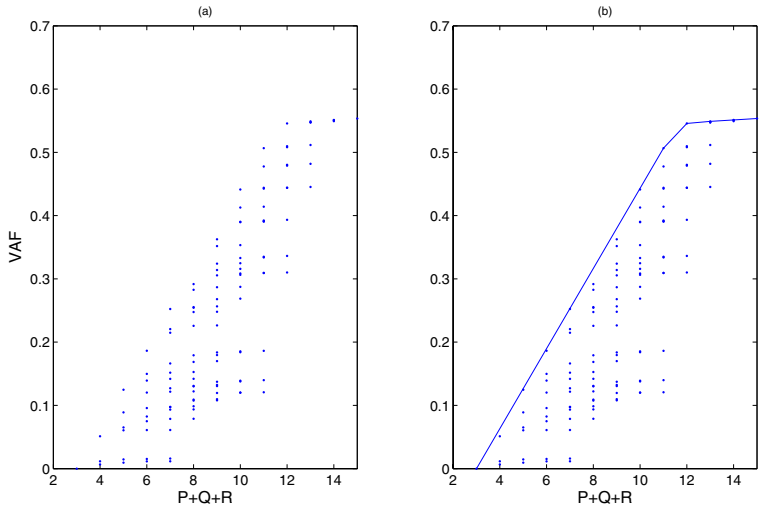


Figure 1. Graphical illustration summarizing the steps in the numerical convex hull heuristic: (a) all solutions are plotted with their VAF-value as a function of $P + Q + R$; (b) the solutions that lie on the upper boundary of the convex hull are connected by a line, one of the solutions with $P + Q + R = 12$ has the highest st-value and will be the final solution selected by this model selection heuristic.

dicating the true underlying complexity in the data. First, we discuss the design of the simulation study and the performance of the four model selection heuristics 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, two 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; and a data matrix **D**, which is **T** perturbed with error. Four design factors were fully crossed on the level of the data generation (matrices **T** and **D**) in this simulation study:

1. the *Size* $I \times J$ of the data matrices at three levels: 40×40, 80×20, 500×50;

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* in the data at two levels: 15, 45%.
4. the *Equality of Cluster Sizes*, at two levels, according to the values of the category probabilities of a multinomial distribution: in the condition *equal*, for both modes the probability of an element being assigned to a cluster is equal for all clusters; in the condition *unequal*, for both modes the category probability for one cluster is $.70$, with the category probabilities for the remaining clusters being equal.

For each cell of the design ten replications were considered, yielding 600 two-mode simulated data sets. In particular, for each combination of the levels of *Size*, *Numbers of Clusters*, *Error* and *Equality of Cluster Sizes*, 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

$$\mathbf{D} = \mathbf{AWB}' + \epsilon\mathbf{E} = \mathbf{T} + \epsilon\mathbf{E}\,, \tag{14}$$

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 $||\mathbf{T}|| = ||\mathbf{E}||$. Note that the uniform distribution for the entries of **W** was chosen as a formalization of a noninformative generating mechanism and that the Gaussian distribution is a standard choice for error distributions.

Each of these 600 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 heuristics were applied to the obtained 25 solutions.

4.1.2 Results

Figure 2 displays the proportion of data sets for which each of the four model selection heuristics indicate the correct underlying numbers of clusters for each of the levels of the four factors in the design (*Size*, *Numbers of Clusters*, *Error* and *Equality of Cluster Sizes*). Clearly, the numerical convex hull procedure and DIFFIT outperform the Extended Calinski-Harabasz and Silhouette indices (with overall hit percentages of 58 and 57 versus 19 and 28, respectively).

In order to investigate the effect of the four design factors on the performance of each of the four model selection heuristics, we analyzed these proportions by means of a repeated multivariate measures analysis

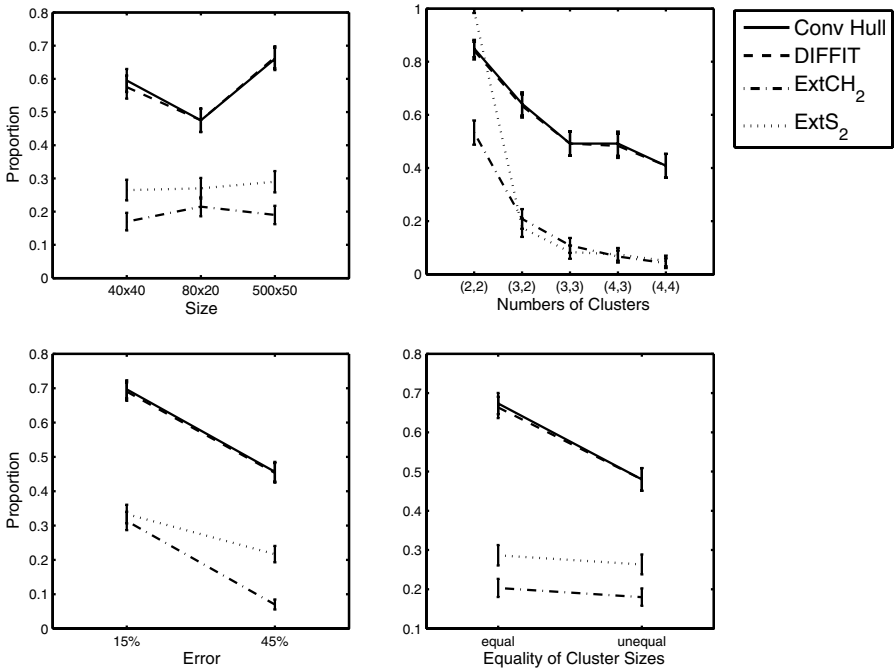


Figure 2. Two-mode simulation study: Proportion of correct model selections by each of the four model selection heuristics under study. Results are split up according to the levels of (a) *Size*, (b) *Numbers of Clusters*, (c) *Error*, (d) *Equality of Cluster Sizes*. Error bars display confidence intervals of two standard error units in length.

of variance (RMANOVA), using only main effects and first-order interactions of model selection heuristic and each of the four design factors. Table 1 displays the effect sizes (partial η^2, computed by dividing SS_{effect} by $SS_{effect} + SS_{error}$) for each of the effects. As each of the interaction effects explain a fairly large amount of variance in the performance of the four model selection heuristics, we will immediately discuss our results in light of these interactions. The largest effect size is observed for the interaction of *model selection heuristic* and *Numbers of Clusters* (partial $\eta^2 = .48$). It can be seen in Figure 2 (b) that this 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 *Numbers of Clusters* becomes larger than $(2, 2)$. Two remarks are important here: First, as discussed in section 3.2, the Extended Silhouette index does not allow the selection of solutions 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

Table 1. Effect sizes calculated for both simulation studies for all main and first-order interaction effects of the four design factors and *model selection heuristic*.

Effect	Effect Size (partial η^2) in two-mode simulation study	Effect Size (partial η^2) in three-mode simulation study
Size	.12	.07
Numbers of Clusters	.80	.72
Error	.50	.66
Equality of Cluster Sizes	.21	.54
model selection heuristic	.76	.91
model selection heuristic × *Size*	.15	.01
model selection heuristic × *Numbers of Clusters*	.48	.54
model selection heuristic × *Error*	.07	.25
model selection heuristic × *Equality of Cluster Sizes*	.15	.01

selection of the least complex model possible for this index. Second, closer inspection of the solutions selected by this index revealed that for only 3% of all data sets a solution was chosen with a larger number of clusters, for any of the two modes, than the true underlying solution. Taking into account these two remarks, it is not surprising that this index performs well when the true numbers of clusters equals $(2, 2)$. The next important interaction effects observed are those of *model selection heuristic* and *Size* and *model selection heuristic* and *Equality of Cluster Sizes* (partial η^2 in both cases being equal to .15). From Figure 2 (a) we see that the first effect is mainly due to a lesser performance of the two best model selection heuristics when *Size* equals 80×20. As the performance of these two model selection heuristics is substantially better in the condition where *Size* equals 500×50, this suggests that this lesser performance can be explained by the fact that one mode contains only a small number of elements (i.e., 20), rather than by a large ratio of number of objects to number of attributes (Otherwise, in comparison with *Size* equal to 80×20, one would expect a worse performance for *Size* equal to 500×50). With respect to the interaction of model selection heuristic and *Equality of Cluster Sizes* we see in Figure 2 (d) that performance decreases when unequal cluster sizes are present, especially for the two best model selection heuristics. Finally, from Figure 2 (c) it can be seen that performance of all four model selection heuristics decreases when the amount of error in the data increases. However, the interaction of *model selection heuristic* and *Error* does not seem to be substantial, as is supported by a smaller value for partial η^2 (i.e., .07).

We further took a closer look at only those cases where the two best model selection heuristics selected an incorrect solution. It turns out that both the numerical convex hull heuristic and DIFFIT underestimate the num-

bers of clusters in 94% of the cases. This suggests that, in case one of these model selection heuristics selects an incorrect solution, there is a tendency to underestimate the numbers of clusters in a data set.

4.2 Three-mode Partitioning

4.2.1 Design

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

1. the *Size* $I \times J \times K$ of the data arrays at four levels: $20 \times 20 \times 20$, $30 \times 30 \times 9$, $80 \times 10 \times 10$, $300 \times 20 \times 10$;
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* in the data at two levels: 15, 45%.
4. the *Equality of Cluster Sizes*, at two levels, according to the values of the category probabilities of a multinomial distribution: in the condition *equal*, for all three modes the probability of an element being assigned to a cluster is equal for all clusters; in the condition *unequal*, for all three modes modes the category probability for one cluster is .70, with the category probabilities for the remaining clusters being equal.

By generating ten replicates in each cell of the design by $\mathbf{D}_A = \mathbf{A}\mathbf{W}_A(\mathbf{C}' \otimes \mathbf{B}') + \epsilon\mathbf{E}$, a total of 800 data sets are considered in this simulation study. Each of these 800 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 heuristics was applied to the resulting 125 solutions.

4.2.2 Results

Figure 3 displays the proportion of data sets for which each of the four model selection heuristics indicate the correct underlying numbers of clusters for each of the levels of the four factors in the design. Clearly, as in the two-mode case, DIFFIT and the numerical convex hull heuristic are superior to the Extended Calinski-Harabasz and Silhouette indices (with overall hit percentages of 91 and 90 versus 24 and 36, respectively). Note that, compared to the two-mode case, performance of the model selection heuristics is better, especially for DIFFIT and the numerical convex hull.

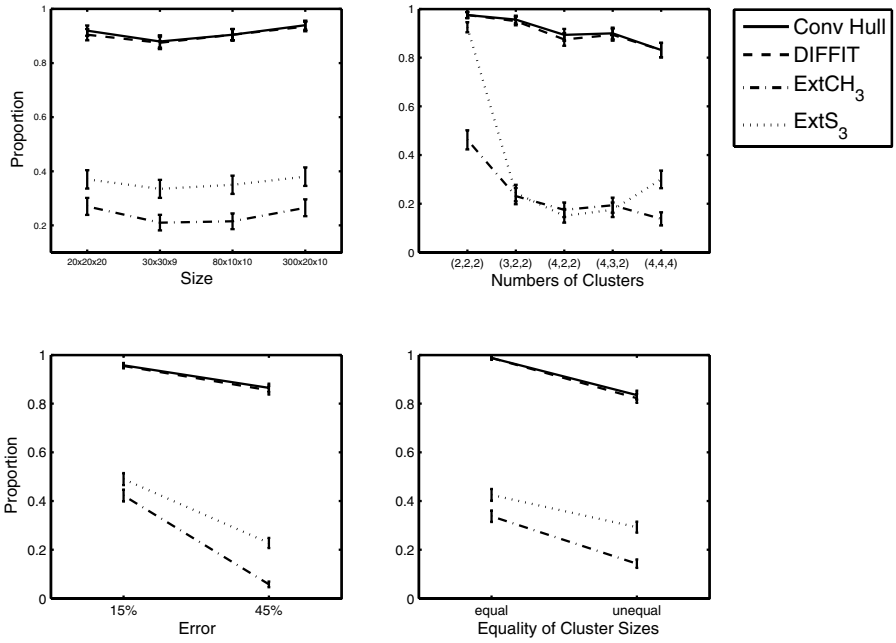


Figure 3. Three-mode simulation study: Proportion of correct model selections by each of the four model selection heuristics under study. Results are split up according to the levels of (a) *Size*, (b) *Numbers of Clusters*, (c) *Error*, (d) *Equality of Cluster Sizes*. Error bars display confidence intervals of two standard error units in length.

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 heuristic and the four design factors. Effect sizes (partial η^2) for each of the effects in the three-mode simulation study are displayed in Table 1. Results are similar to the two-mode case with *model selection heuristic* being the most important factor (partial $\eta^2 = .91$) in explaining the variability in the proportion of data sets for which a correct model was indicated. However, some deviations in effect sizes from the two-mode case are observed. First, the effects of the interaction of *model selection heuristic* and *Size* and of *model selection heuristic* and *Equality of Cluster Sizes* are not as important as in the two-mode case (partial η^2 in both cases being equal to $.01$). Second, the size of the effect of the interaction of *model selection heuristic* and *Error* is now clearly larger (partial $\eta^2 = .25$). Third, the main effect of *Equality of Cluster Sizes* has gained in importance (partial $\eta^2 = .54$). Summarizing these results, we see that performance decreases when the number of underlying clusters increases (see Figure 3 (b)),

especially for the Extended Silhouette index (A similar explanation as for the two-mode simulation can be given here); performance decreases when the amount of error in the data increases (see Figure 3 (c)), but this is less so for DIFFIT and the numerical convex hull method; and all four model selection heuristics perform worse when unequal cluster sizes are present (see Figure 3 (d)).

5. Application to a Collection of Mixed Empirical-artificial Data Sets

One may wonder to what extent the results found in the simulation studies above (in which the artificial data sets were generated in a standard way), 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) $\times$ 20(scales) $\times$ 38(participants) data array $\underline{\mathbf{D}}$ by centering 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 $\underline{\mathbf{M}}$ according to (3), (b) obtaining the residual array $\underline{\mathbf{R}}$ by subtracting $\underline{\mathbf{M}}$ from $\underline{\mathbf{D}}$, 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 $\underline{\mathbf{R}}$, to the model array $\underline{\mathbf{M}}$. 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 i.i.d. 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 perform best (in this case perfectly). The performance of the extended Calinski-Harabasz index is even worse (the correct model was never chosen) than in the conditions that were most comparable in terms of size (i.e., $20 \times 20 \times 20$ and $30 \times 30 \times 9$) in the simulation study with artificial data. This finding suggests that the performance of this index is not robust to departures from data as generated in the simulation studies. Closer examination reveals that, for all three modes it always selected a solution with too few clusters. The extended Silhouette index on the other hand seems to perform somewhat better than expected from our simulation studies. 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. Based on these findings, and on the findings reported in the two simulation studies, we recommend researchers applying multi-mode partitioning models to select DIFFIT or the numerical convex hull as the model selection heuristic of their choice.

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 heuristics ranged from satisfactory (in the two-mode case) to very good (in the three-mode case). The difference between the model selection heuristics further was found to be the most important factor in explaining the performance in indicating the correct numbers of clusters. In both the two- and three-mode partitioning cases, 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. Finally, for both the two- and three-mode partitioning cases it can be concluded that performance of all four heuristics decreases when the amount of error increases and when un-

equal cluster sizes are present, regardless of whether the effect of the interaction of *model selection heuristic* and *Error* and of *model selection heuristic* and *Equality of Cluster Sizes* turned out to be highly important or not in the analyses.

The multi-way model selection heuristics 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 heuristics 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 heuristics.

The other two methods considered in this paper are merely straightforward extensions of model selection heuristics 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., and SCHADER, M. (1997), "Two-Mode Overlapping Clustering with Applications to Simultaneous Benefit Segmentation and Market Structuring", in *Classification and Knowledge Organization*, Eds. R. Klar and O. Opitz, Berlin: Springer, pp. 557–566.

CALINSKI, R.B. and HARABASZ, J. (1974), "A Dendrite Method for Cluster Analysis", *Communications in Statistics, 3*, 1–27.

CASTILLO, W. and TREJOS, J. (2002), "Two-mode Partitioning: Review of Methods and Application of Tabu Search", in *Classification, Clustering, and Related Topics. Recent Advances and Applications. Studies in Classification, Data Analysis, and Knowledge Organization*, Eds. K. Jajuga, A. Sokolowski, and H.H. Bock, Heidelberg: Springer–Verlag, pp. 43–51.

CEULEMANS, E. and KIERS, H.A.L. (2006), "Selecting Among Three-mode 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. and VAN MECHELEN, I. (2005), "Hierarchical Classes Models for Three-way Three-mode Binary Data: Interrelations and Model Selection", *Psychometrika, 70*, 461–480.

EVERITT, B.S. (1993), *Cluster Analysis*, Arnold: London.

GAUL, W. and SCHADER, M. (1996), "A New Algorithm for Two-mode Clustering", in *Classification and Knowledge Organization*, Eds. H. Bock and W. Polasek, Berlin: Springer, pp. 15–23.

HARTIGAN, J.A. (1975), *Clustering Algorithms*, New York: Wiley.

KAUFFMAN, L. and 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*, Paper presented at COMPSTAT 2004, Prague.

KROONENBERG, P.M. and OORT, F.J. (2003), "Three-mode Analysis of Multimode Covariance Matrices", *British Journal of Mathematical and Statistical Psychology, 56*, 305-336.

KROONENBERG, P.M. and VAN DER VOORT, T.H.A. (1987), "Multiplicative Decomposition of Interactions for Judgements of Realism of Television Films", *Kwantitatieve Methoden, 8*, 117–144.

MACQUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1"*, Eds. L.M. LeCam and J. Neyman, Berkeley: University of California Press, pp. 281–297.

MILLIGAN, G.W. and COOPER, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika, 50*, 159–179.

MURAKAMI, T. and KROONENBERG, P.M. (2003), "Three-mode Models and Individual Differences in Semantic Differential Data", *Multivariate Behavioral Research, 38*, 247–283.

ROCCI, R. and VICHI, M. (2008), "Two-mode Multi-Partitioning", *Computational Statistics and Data Analysis, 52*, 1984–2003.

SCHEPERS, J., VAN MECHELEN, I., and CEULEMANS, E. (2006), "Three-mode Partitioning", *Computational Statistics & Data Analysis, 51*, 1623–1642.

SUGAR, C.A. and JAMES, G.M. (2003), "Identifying Groups in Data: An Information-Theoretic Approach", *Journal of the American Statistical Association, 98*, 750–763.

TIMMERMAN, M.E. and KIERS, H.A.L. (2000), "Three-mode Principal Components Analysis: Choosing the Numbers of Components and Sensitivity to Local Optima", *British Journal of Mathematical and Statistical Psychology, 53*, 1–16.

TIBSHIRANI, R., WALTHER, G., and 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., and CASTILLO, W. (2005), "Global Optimization Strategies for Two-Mode Clustering", Econometric Institute Report EI 2005-33, Erasmus School of Economics, Erasmus University Rotterdam.

VICHI, M. (2002), "Double k-means Clustering for Simultaneous Classification of Objects and Variables", in *Advances in Classification and Data Analysis*, Eds. S. Borra, R. Rocci, and M. Schader, Heidelberg: Springer, pp. 43–51.

WANSBEEK, T. and VERHEES, J. (1989), "Models for Multidimensional Matrices in Econometrics and Psychometrics", in *Multiway Data Analysis*, Eds. R. Coppi and S. Bolasco, Amsterdam: Elsevier, pp. 543–552.