TECHNICAL
REPORT

0480

THREE-MODE PARTITIONING

SCHEPERS,J. and I. VAN MECHELEN

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

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

Three-mode partitioning

Jan Schepers

Iven Van Mechelen

Katholieke Universiteit Leuven

Author Notes:

Correspondence concerning this paper should be addressed to Jan Schepers, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium. Email: Jan.Schepers@psy.kuleuven.ac.be.

Running head: Three-mode partitioning

Abstract

The three-mode partitioning model is a clustering model for three-way three-mode data sets that implies a simultaneous partitioning of all three modes of the data set. In the associated data analysis, a data array $\underline{\mathbf{D}}$ is approximated by a model array $\underline{\mathbf{M}}$ that can be represented by a three-mode partitioning model of a prespecified rank (P, Q, R), minimizing a least-squares loss function in terms of differences between $\underline{\mathbf{D}}$ and $\underline{\mathbf{M}}$. The present paper is concerned with the estimation of the model. First, a framework for alternating least squares methods is described for dealing with the estimation problem in question. Next, a number of algorithms is proposed within this framework. An extensive simulation study is reported in which these algorithms are evaluated and compared according to sensitivity to local optima and recovery of truth underlying the data. When applying the algorithms to a collection of four empirical data sets, the ordering of the algorithms with respect to performance in finding the optimal solution appears to change as compared to the results obtained from the simulation study. This finding is attributed to violations of the implicit stochastic model underlying the simulation study. Support for the latter attribution is found in a second simulation study.

1 Introduction

N-way data in general offer a wealth of information to the data analyst. However, this information is usually very complex and hard to grasp. A way out for this can be to reduce one or more of the involved modes to a small number of classes or dimensions. In this paper, we are concerned with a simultaneous reduction of the three modes of a three-way three-mode data set. More in particular, we will focus on the most simple case: the simultaneous partitioning of all three modes involved in the data. Recently, the Three-mode Partitioning model was proposed independently by several authors (Rocci & Vichi, 2003; Kiers, 2004; Schepers & Van Mechelen, 2004). The model is in fact a direct generalization of the simultaneous two-mode partitioning models proposed by Gaul and Schader (1996), Baier, Gaul and Schader (1997), Govaert (1995) and Vichi (2002). For a comprehensive overview of two-mode clustering methods, see Van Mechelen, Bock and De Boeck (2004). A problem that is still left unresolved is the estimation of the three-mode partitioning model. More in particular, a number of k-means type algorithms have been proposed by the different authors of the model, but so far, the performance of these algorithms is not yet clear, let alone, what is the preferred method. One of the reasons to be careful in this regard is the well-known fact that k-means type algorithms suffer from local minima problems (Selim & Ismail, 1984; Steinley, 2003). Considering the increased complexity when clustering more than a single mode, there seems to be an obvious need to assess the performance of algorithms for the three-mode partitioning situation. The present paper will examine the performance of several three-mode k-means type algorithms by means of a simulation study and by means of a test on four empirical data sets. The remainder of the paper is organized as follows: Section 2 recapitulates the Three-Mode Partitioning model. Section 3 presents a framework for alternating least squares methods for estimating the model. In Section 4, a simulation study is presented in which several algorithms are evaluated in terms of their capability of minimizing the loss function and in terms of their capability to recover the truth underlying the data. In Section 5, we report the results of the application of the same algorithms to four empirical data sets. Unexpectedly, the latter results will appear not to be in line with those of the simulation study. An explanation of this finding will be looked for by means of parametric bootstrap tests. This explanation will further be tested in a second simulation study that will be presented in Section 6. Section 7 will present a few concluding remarks.

2 Three-Mode Partitioning

2.1 Data

A three-way data set defines a mapping from the Cartesian product of three sets of entities to some value set (say the set of reals $\mathbb{R}$). If the three sets of the Cartesian product are all distinct, the data set is referred to as three-way three-mode (Carroll & Arabie, 1980). In many areas of sciences, this type of data set often occurs. Examples include data in personality psychology pertaining to the intensity of different behaviors as elicited by various situations for different persons, and data in marketing research pertaining to the perceived usefulness of different products for different goals as judged by diverse age groups. In the remainder of this paper, we will refer to the horizontal slices of three-way three-mode data as objects, to the lateral slices as attributes and to the frontal slices as sources. A three-mode partitioning model further requires array-conditionality, implying that any two data entries are comparable. Furthermore, the particular set of partitioning models that we will consider in the present paper also require three-way three-mode problems that imply the aim of reconstructing the actual data values (rather than of summarizing dependency information, as implied by a three-way three-mode contingency table, or of summarizing interaction information, as implied by three-way three-mode data on the value of a criterion variable associated with all combinations of values of three categorical predictor variables). In the following, we will denote the data points by d_{ijk}, referring to the value of the i-th object on the j-th attribute according to the k-th source.

2.2 Models

A three-mode partitioning implies a decomposition of an $I \times J \times K$ real-valued model or reconstructed data array $\underline{\mathbf{M}}$ into an $I \times P$ object partition matrix $\mathbf{A}$, a $J \times Q$ attribute partition matrix $\mathbf{B}$, a $K \times R$ source partition matrix $\mathbf{C}$ and a $P \times Q \times R$ real-valued array $\underline{\mathbf{W}}$, with (P, Q, R) being the rank of the model. More in particular the model array can be written as:

$$m_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} w_{pqr} \quad \forall i, j, k \tag{1}$$

where a_{ip}, b_{jq} and c_{kr} indicate whether or not object i, attribute j and source k belong to cluster p, q and r, respectively, and where w_{pqr} is a real-valued number representing the strength of the relation between clusters p, q and r. The rows of the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are restricted to sum to 1 and no empty clusters (zero columns) are allowed so as to represent proper partitioning structures. Equation (1) implies that m_{ijk} equals w_{pqr} iff object i, attribute j and source k belong to object cluster p, attribute cluster q and source cluster r, respectively. Of course, small deviations between the actual and reconstructed data points are to be expected in practice. In order to link the model to a given data set, an error term e_{ijk} is therefore assumed,

$$d_{ijk} = m_{ijk} + e_{ijk} \quad \forall i, j, k \tag{2}$$

In absence of further assumptions on e_{ijk}, (1) can be considered to denote a deterministic model. Optionally, however, a stochastic version of (1) can be obtained by making additional assumptions about the distribution of the error component in (2), e.g. that the e_{ijk} are iid normally distributed with zero mean, yielding:

$$d_{ijk} \overset{iid}{\sim} N(m_{ijk}, \sigma^2_{ijk}) \tag{3}$$

where the mean m_{ijk} is defined as in (1) and the variance σ^2_{ijk} is the variance of the error component. Typically, further restrictions may be put on the error distribution. For example, all σ^2_{ijk} can be assumed to be equal to a single σ^2 (homoscedastic case). A less restrictive assumption could read that σ^2 depends on the data cluster (p, q, r) to which the data point (i, j, k) is assigned. Note that in all cases as considered in the present paper the entries of the **A**, **B** and **C** matrices are considered fixed constants (rather than as realizations of latent multinomially distributed variables), implying that the stochastic models under study can be considered so-called fixed-partition models (Bock, 1996).

2.3 Graphical Representation

A three-mode partitioning model can be given a graphical representation in the form of a heat map that visualizes the elements of the core array $\underline{\mathbf{W}}$ in such a way that dark colors represent high values and light colors low ones. An example of such a heat map for a three-mode partitioning of a hypothetical data set is given in Figure 1. This type of visualization nicely illustrates the data compression that is achieved by a three-mode partitioning: whereas the original data array $\underline{\mathbf{D}}$ consists of $I \times J \times K$ elements, the heat map of the core $\underline{\mathbf{W}}$ includes only $P \times Q \times R$ different cells.

insert Figure 1 here

In an alternative heat map representation, one may display the actual rather than the reconstructed data values, after permuting the entities of all three modes in such a way that objects (resp. attributes, sources) assigned to the same cluster are positioned adjacent to one another. The advantage of the latter type of visualization is that, apart from the partitioning structure, it also allows one to capture how well the model explains the data. Note that this type of visualization fits within a non-destructive data analysis approach (Murtagh, 1989). Figure 2 illustrates this type of representation for the same hypothetical example as in Figure 1.

insert Figure 2 here

2.4 Data Analysis

In order to fit the deterministic model in (1) to a given data set $\underline{\mathbf{D}}$, the following loss function is minimized:

$$f(\underline{\mathbf{M}}) \quad = \quad \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(d_{ijk} - m_{ijk})^2. \tag{4}$$

where d_{ijk} is the data value for object i, attribute j and source k, and m_{ijk} is defined as in (1). Expression (4) shows that the model is fitted in terms of a least squares loss function implying that one will try to find data clusters that are as homogeneous as possible in the least squares sense. Constructing procedures for jointly estimating the partition matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ in such a way that (4) is minimized is not straightforward (see Section 3 below). However, the estimation of $\underline{\mathbf{W}}$, conditionally on $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ follows directly from (1) and (4):

$$w_{pqr} \quad = \quad \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} a_{ip}b_{jq}c_{kr}d_{ijk}}{(\sum_{i=1}^{I} a_{ip})(\sum_{j=1}^{J} b_{jq})(\sum_{k=1}^{K} c_{kr})} \qquad \forall p, q, r \tag{5}$$

implying that the best possible estimate of w_{pqr} is the mean of all the data points for which it holds that the corresponding objects, attributes and sources belong to clusters p, q and r, respectively.

The stochastic version of the three-mode partitioning model implies to find a triple of partitions: $\mathcal{A} = \{\mathcal{A}_1, ..., \mathcal{A}_P\}$, $\mathcal{B} = \{\mathcal{B}_1, ..., \mathcal{B}_Q\}$ and $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_R\}$, the array $\underline{\mathbf{W}}$ and the parameters of the error component distribution such that the likelihood

$$f(\underline{\mathbf{D}}; \theta, \mathcal{A}, \mathcal{B}, \mathcal{C}) \quad = \quad \prod_{p}^{P}\prod_{q}^{Q}\prod_{r}^{R}\prod_{i\in\mathcal{A}_p}\prod_{j\in\mathcal{B}_q}\prod_{k\in\mathcal{C}_r}\psi(d_{ijk}; \alpha_{pqr}) \tag{6}$$

is maximized. The function ψ is the density of the observation d_{ijk}, e.g., the normal density as in (3) (In the latter case, the parameter vector α_{pqr} consists of the mean and variance of the data cluster to which d_{ijk} is assigned).

Taking the natural logarithm, (6) becomes:

$$L(\theta, \mathcal{A}, \mathcal{B}, \mathcal{C}) \quad = \quad \log(f(\underline{\mathbf{D}}; \theta, \mathcal{A}, \mathcal{B}, \mathcal{C}))$$

$$= \sum_{p}^{P}\sum_{q}^{Q}\sum_{r}^{R}\sum_{i\in\mathcal{A}_p}\sum_{j\in\mathcal{B}_q}\sum_{k\in\mathcal{C}_r}\log(\psi(d_{ijk};\alpha_{pqr})) \qquad (7)$$

The estimation problem as outlined above fits within the classification likelihood approach to clustering (see e.g. John, 1970; Bryant & Williamson, 1978; McLachlan, 1982; Celeux & Govaert, 1992; Banfield & Raftery, 1993; Govaert & Nadif, 2003). A distinctive feature from the perhaps better known mixture modeling approach to clustering, is that in the classification approach, the partition class memberships show up in the likelihood and are treated as a set of parameters that are to be estimated. For the one-mode partitioning situation, Scott and Symons (1971) pointed out that several common deterministic clustering procedures, such as one-mode k-means, can be shown to be special cases of a classification likelihood approach. An analogous relation holds for the three-mode partitioning case: minimizing the loss function in (4) is in fact equivalent to maximizing (7) when in the latter case ψ is assumed normally distributed with parameters (m_{ijk}, σ^2) with m_{ijk} being defined as in (1). In the remainder of this paper, we will focus on the specific situation where minimizing (4) is equivalent to maximizing (7).

2.5 Uniqueness

The solution that minimizes loss function (4) is in general unique upon permutations of the partitions **A**, **B** and **C**.

Proof:

It is convenient to write the model (1) in matrix notation: $M_A = AW_A(C' \otimes B')$, where M_A is the matricized $I \times JK$ form of the model array $\underline{\mathbf{M}}$, W_A is the matricized $P \times QR$ form of the core $\underline{\mathbf{W}}$, and $\otimes$ represents the Kronecker product. The partition matrix A can be multiplied by any non-singular matrix T without having an effect on the loss function (4), as long as the inverse transformation is applied to the core W_A: $M_A = (AT)(T^{-1}W_A)(C' \otimes B')$. However, in order to retain a partitioning model, AT must be a partition matrix. Thus,

- T must be a binary matrix in order for AT to be binary
- T can have only a single 1 per row, because otherwise AT will have rows with more than a single 1, violating the partitioning structure
- T can also have only a single 1 per column, because otherwise, taking into account that T can have only a single 1 per row, T would be singular

the only matrices that satisfy the previous three requirements are permutations matrices.

3 Framework for ALS Procedures

In the present paper, we will focuss on ALS (Alternating Least Squares) methods to estimate the three-mode partitioning model. In this section, a framework for the latter methods is described. Often used in data-analysis methods, ALS-procedures estimate one set of parameters, keeping another set of parameters fixed. By alternatingly switching between the different parameter sets (estimating one set keeping the other set fixed, and vice versa), the loss function decreases (or at least stays the same) at every alternate step. Since the loss function (4) is bounded below, the algorithm necessarily converges. Note however that the ALS-procedure is a heuristic method and thus may arrive at local optima of the solution space.

In constructing an ALS algorithm for estimating a three-mode partitioning model, a choice has to be made with respect to three parts. First, one needs an initial value for each parameter that is to be estimated. This implies that an initial estimate is needed for each of the modes, and for the core. Different procedures to obtain these initial values will be discussed in Subsection 3.1. Second, the general updating scheme within the alternating sequence of obtaining conditional estimates of each parameter set may take different forms. This topic will be discussed in Subsection 3.2. Finally, in the progress of estimating the three-mode partitioning model, given a data set, empty clusters might turn up. As already mentioned in Subsection 2.2, this is not allowed because empty columns in the matrices **A**, **B** and **C** imply improper partitions. A variety of strategies can be thought of as to how to deal with these empty clusters. This topic will be discussed in Subsection 3.3.

3.1 Initialization of the Partition Matrices

Because all alternating least squares algorithms start with an initial estimate of the unknown parameters, the question arises as to what kind of initialization produces a high probability of the final estimates equalling the global minimum of the loss function (4). In a general context of clustering problems, a smart choice for the initial estimates, as opposed to starting from a randomly drawn position in the solution space, is usually thought to provide better results with respect to the final solution returned by the clustering algorithm. The general idea is that a smart initial estimate usually lies close to the globally optimal solution and therefore the ALS-algorithm is, while proceeding on its way to the optimal solution, less likely to get stuck in locally optimal solutions. The idea of using a smart (rational) start has already received attention in the one-mode partitioning literature and simulation results indeed suggest that better k-means solutions can be obtained by using a rational start (Milligan, 1980). As a consequence, most statistical packages for k-means clustering provide the user the ability to use a rational starting procedure. It must be noted however that Steinley (2003), in a more thorough investigation, examined the local minima problem for the k-means clustering method, as implemented in 3 well-known commercial software packages (SPSS, SYSTAT and SAS), which all use a single rational starting configuration. By comparing the results provided by these implementations with the results of using a multistart procedure (with every single initial estimates representing a randomly drawn position in the solution space), he concluded that the methods implemented in the commercial packages are very likely to end up in local optima. The author suggests that a procedure where the k-means method is run from several alternative starting points is more likely to provide the optimal estimates, and thus more appropriate when faced with the problem of trying to find an optimal one-mode partitioning, given a data set. The results of this study therefore suggest that, when it comes down to finding the optimal k-means clustering, the "quality of the starts" is less important than "quantity of the starts".

In this paper, we will try to deal with the conclusions from the previous research on one-mode k-means clustering by incorporating different kinds of rational starts, a random multi-start procedure analogous to the one used by Steinley (2003), and also a rational multi-start procedure. First, we will highlight the different types of rational starts considered in the present paper. Then, the random multi-start procedure is explained and finally the construction of multiple rational starts will be elaborated.

Three types of rational starts for the estimation of the three-mode partitioning model are considered in the present paper. The first procedure to obtain a smart start is to perform separate one-mode k-means analyses on the (appropriately) matricized forms of the three-way data array and use their outcomes as initial estimates of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively. For example, when a rational choice for the initial matrix $\mathbf{A}$ is desired, one first "flattens out" the three-mode data array $\underline{\mathbf{D}}$ to a data matrix $\mathbf{D}$ with dimensions $I \times JK$, and then searches for the optimal one-mode k-means partitioning of the objects, based on their patterns on the set of variables (which consists of all elements of the Cartesian product of attributes and sources). Note that one faces the same problem of choosing an initial estimate for the separate one-mode partitioning analyses. For a discussion of the possibilities and some results with respect to the one-mode partitioning situation, the reader is referred to Milligan (1980). We will further refer to this rational start as *Independent k-means Start.*

The second type of rational start for the estimation of the three-mode partitioning model was proposed by Kiers (2004) who suggested that a useful starting configuration can be obtained by performing a Tucker3 (see e.g. Kroonenberg & De Leeuw, 1980) analysis on the data set. This leads to columnwise orthonormal component matrices for the three modes. Then, by performing a varimax rotation of all three component matrices and finally, after multiplying columns having negative sums by -1, a partition matrix is obtained by setting all rowwise highest elements to 1, and all other elements to 0. We will further refer to this type of rational start as *Tucker3 Start*.

The final type of rational start considered in this paper was proposed by Rocci and Vichi (2004). Their sequential k-means procedure first searches for example for an optimal one-mode k-means partitioning of the objects. Then, an optimal partition of the attributes is obtained using the matrix of centroids from the first clustering procedure, after being weighted by the cardinalities of the object clusters, as the input data matrix. Finally, an optimal partition of the sources is found using the centroid matrix from the second clustering procedure, weighted by the cardinalities of object and attribute clusters. Note that, when searching for this type of initial triplet of partitions, there are 6 different ways in which the sequential procedure can be applied. In this paper, we only use the sequence as described above. We will further refer to this type of rational start as *Sequential k-means Start*.

If the starting configuration is not obtained by a rational procedure but instead one leaves the partitioning structure of the initial estimates to be determined completely by chance, one can proceed in the following way: randomly assign each object (resp. attribute, source) to a cluster p (resp. q, r), with the probability of being assigned to a cluster p (resp. q, r) equal for all $p = 1, ..., P$ (resp. $q = 1, ..., Q$, $r = 1, ..., R$). An advantage of this procedure is that a large amount of initial starting points can be generated on which subsequently an ALS algorithm can be run from each separate start. As already described, Steinley (2003) showed that this reduces the problem of getting stuck in local minima. We will further refer to this type of start as *Random Multi-Start*.

Finally, since rational starting procedures only provide one single initial starting point, we use the following procedure to obtain multiple alternative rational starting configurations: given a rational start ($\mathbf{A}^R$, $\mathbf{B}^R$, $\mathbf{C}^R$) for each of the partition matrices, randomly choose a proportion of 20% of the objects (resp. sources, attributes), identify to which cluster these elements are assigned (according to the rational start), and reassign them to one of the other clusters corresponding to that mode. This produces a slightly perturbed rational start ($\mathbf{A}^{pR}$, $\mathbf{B}^{pR}$, $\mathbf{C}^{pR}$). By repeating this procedure N times, one obtains a set of N alternative starting points $\{(\mathbf{A}_1^{pR}$, $\mathbf{B}_1^{pR}$, $\mathbf{C}_1^{pR})$,..., $(\mathbf{A}_N^{pR}$, $\mathbf{B}_N^{pR}$, $\mathbf{C}_N^{pR})\}$, all representing random perturbations of the single initialization obtained by applying a rational procedure. In the remainder, we will denote the set of N perturbed rational starts generated from the set ($\mathbf{A}^R$, $\mathbf{B}^R$, $\mathbf{C}^R$) as obtained by *Independent k-means Start* as *Independent k-means Multi-Start*. When the N alternative starts were generated from the set ($\mathbf{A}^R$, $\mathbf{B}^R$, $\mathbf{C}^R$) as obtained by *Tucker3 Start*, we will use the term *Tucker3 Multi-Start* and finally, in case of N starts resulting from *Sequential k-means Start*, we will refer to this type of start as *Sequential k-means Multi-Start*. Note that the use of rational multi-starts was already proposed by Ceulemans (??) in a different clustering context. The author found a significant increase in performance in estimating the HICLAS model (De Boeck & Rosenberg, 1988) when a rational multi-start procedure was used.

3.2 ALS-scheme

Assuming initial estimates for the partition matrices, a further choice has to be made concerning the estimation of the three-mode partitioning model. The general ALS-scheme for estimating a three-mode partitioning will be described in this subsection.

Four distinct sets of parameters can be distinguished in this optimization problem: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{W}}$. The sequence in which one of these four sets is estimated, keeping the other parameter sets fixed, defines the nature of the ALS-scheme. We distinguish three types of ALS-scheme in this paper.

If no distinction is made between the respective natures of one of these four sets of parameters, it is normal to sequentially estimate one set, keeping the other sets fixed, and passing through each member of the four parameter sets in a one by one fashion. An example of this type of ALS-scheme is to first estimate $\mathbf{A}$ (holding $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{W}}$ fixed), then estimating $\mathbf{B}$ (holding $\mathbf{A}$, $\mathbf{C}$ and $\underline{\mathbf{W}}$ fixed), then estimating $\mathbf{C}$ (holding $\mathbf{A}$, $\mathbf{B}$ and $\underline{\mathbf{W}}$ fixed), and finally estimating $\underline{\mathbf{W}}$ (holding $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ fixed). The latter procedure can be described by the following ALS-chain ($\mathbf{ABC}\underline{\mathbf{W}}\mathbf{AB}$...). In total, 4! of these chains are possible by taking all permutations of this sequence. In the present paper however, we will take the specific sequence as described above as the representative instantiation of this particular ALS-scheme.

It is clear that the nature of the parameters, from an optimization perspective, is not the same for every single parameter set. More in particular, the cells of $\underline{\mathbf{W}}$ may take any value from the set of reals $\mathbb{R}$, while the cells in $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are constrained to the binary set. Moreover, the optimal value for $\underline{\mathbf{W}}$ (keeping $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ fixed) is, as described in Section 2.4, available in closed form by formula (5). Similar to what is usually called the centroid matrix in one-mode k-means, $\underline{\mathbf{W}}$ can be considered to represent the cluster centers of the three-mode partitioning model. Therefore, would it not be natural to update these centers whenever a difference in clustering structure is observed? Following this line of reasoning, the second ALS-scheme that we consider in this paper will again estimate each of the four parameter sets keeping the other three parameter sets fixed, but as opposed to ALS-chain ($\mathbf{ABC}\underline{\mathbf{W}}\mathbf{AB}$...), $\underline{\mathbf{W}}$ is reestimated after each estimation of one of the partition matrices. An example of this second type of ALS-scheme is to first estimate $\mathbf{A}$ (holding $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{W}}$ fixed), then estimating $\underline{\mathbf{W}}$ (holding $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ fixed), then estimating $\mathbf{B}$ (holding $\mathbf{A}$, $\mathbf{C}$ and $\underline{\mathbf{W}}$ fixed), then again estimating $\underline{\mathbf{W}}$ (holding $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ fixed), and then estimating $\mathbf{C}$ (holding $\mathbf{A}$, $\mathbf{B}$ and $\underline{\mathbf{W}}$ fixed). The latter procedure can be described by the following ALS-chain ($\mathbf{A}\underline{\mathbf{W}}\mathbf{B}\underline{\mathbf{W}}\mathbf{C}\underline{\mathbf{W}}$...). In total, 3! of these chains are possible by taking all permutations of this sequence. In the present paper however, we will take the specific sequence as described here as the representative instantiation of this particular ALS-scheme.

The way in which the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are estimated is necessarily the same in both

ALS-schemes described above. This is due to the particular procedure of keeping the core and the other partition matrices fixed while estimation one partition matrix. Consider the estimation of $\mathbf{A}$, conditionally on $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{W}}$. An analysis of the loss function (4) shows that it satisfies a separability property (Chaturvedi & Carroll, 1994), implying that the contribution of the cluster pattern of object i ($a_{i.}$) to the loss function can be separated from the contribution of the cluster patterns of the other objects. For the conditional estimation of $\mathbf{A}$ (keeping $\mathbf{B}$, $\mathbf{C}$, and $\underline{\mathbf{W}}$ fixed), this property is illustrated by the following decomposition of the loss function (4) in I separate terms:

$$\begin{aligned} & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(d_{ijk} - \sum_{p=1}^{P} a_{ip} \sum_{q=1}^{Q}\sum_{r=1}^{R} b_{jq}c_{kr}w_{pqr})^2 \\ = & \sum_{j=1}^{J}\sum_{k=1}^{K}(d_{1jk} - \sum_{p=1}^{P} a_{1p} \sum_{q=1}^{Q}\sum_{r=1}^{R} b_{jq}c_{kr}w_{pqr})^2 + ... + \\ & \sum_{j=1}^{J}\sum_{k=1}^{K}(d_{Ijk} - \sum_{p=1}^{P} a_{Ip} \sum_{q=1}^{Q}\sum_{r=1}^{R} b_{jq}c_{kr}w_{pqr})^2 \end{aligned} \tag{8}$$

Consequently, the conditionally optimal estimate of $\mathbf{A}$ can be found by independently optimizing the pattern $a_{i.}$ $(i = 1...I)$ in each of the I terms of (8). A conditionally optimal estimate of $\mathbf{A}$ can thus be obtained by assigning each element $i(i = 1...I)$ to the cluster for which it holds that the i'th term in (8) is minimal. With respect to the conditional estimation of $\mathbf{B}$ and $\mathbf{C}$, keeping all other parameter sets fixed, a similar separability property holds when the loss function is decomposed into J, respectively K terms.

Finally, the third ALS-scheme takes the reasoning of the second ALS-scheme to the limits. Instead of reestimating $\underline{\mathbf{W}}$ after all objects (resp. attributes, sources) are assigned to their respective closest cluster centers, one can instead consider to monitor the change in $\underline{\mathbf{W}}$ with every possible change in cluster assignment of each object (resp. attribute, source). Here, instead of fixing the cluster centers, one allows these centers to drift while evaluating every possible cluster assignment. Formally, this comes down to evaluating the loss function where the estimate of $\underline{\mathbf{W}}$ (as computed by expression (5)) is plugged into (1),

$$f'(\underline{\mathbf{M}}) \quad = \quad \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(d_{ijk} - \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}\frac{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J}\sum\limits_{k=1}^{K} a_{ip}b_{jq}c_{kr}d_{ijk}}{\sum\limits_{i=1}^{I} a_{ip}\sum\limits_{j=1}^{J} b_{jq}\sum\limits_{k=1}^{K} c_{kr}})^2 \qquad (9)$$

Any change in **A**, **B** or **C** is now immediately accompanied by its corresponding update in the core. Consider now the estimation of **A**, conditionally on **B** and **C**. Clearly, the separability property as in (8) does not longer hold and it is thus no longer possible to obtain an optimal estimate of **A** by evaluating each row separately. A possible representation of this type of ALS-scheme is $(\mathbf{A}_{\underline{\mathbf{W}}}\mathbf{B}_{\underline{\mathbf{W}}}\mathbf{C}_{\underline{\mathbf{W}}}...)$. From an analytical perspective, the conditional estimation in this ALS-scheme boils down to the one-mode k-means problem because one simultaneously tries to find an optimal least squares clustering of some set together with the cluster centers. From a conditional optimization perspective, this procedure is a lot more difficult than the situation as described for the first two ALS-schemes. In general, finding an optimal clustering for one-mode k-means is only feasible when the number of objects to be clustered is small. However, even in the latter case these methods take a lot of computation time to get to the solution. The same necessarily holds for the conditional optimization within the last ALS-scheme. Note that for this final ALS-scheme there are also 3! possible permutations of the sequence, but we only consider the one represented here in the remainder of this paper.

Three different ALS-schemes were explained in detail in this Section: $(\mathbf{ABC}\underline{\mathbf{W}}\mathbf{AB}...)$, $(\mathbf{A}\underline{\mathbf{W}}\mathbf{B}\underline{\mathbf{W}}\mathbf{C}\underline{\mathbf{W}}...)$ and $(\mathbf{A}_{\underline{\mathbf{W}}}\mathbf{B}_{\underline{\mathbf{W}}}\mathbf{C}_{\underline{\mathbf{W}}}...)$. The difference between these ALS-schemes can be described in the amount of attention that is directed towards the estimation of the cluster centers. The first scheme pays the least attention to this aspect of the estimation problem, the second scheme an intermediate amount of attention and the third scheme the most.

3.3 Empty Clusters

If, after estimating one of the individual partition matrices **A**, **B** and **C**, empty clusters are present, the subsequent calculation of the core becomes problematic because the cluster centers corresponding to the empty clusters are left undefined. Also, from an optimization perspective, the best solution with (P, Q, R) clusters, given a data set, is necessarily better (or equally good) in terms of its corresponding loss function (4) than the best solution with e.g. $(P-1, Q, R)$ clusters (this reflects the case if one of the clusters of the objects mode is empty). Therefore, it seems appropriate to constrain the partition matrices to have all columnwise sums greater than 0 ($\sum_{i=1}^{I} a_{ip} > 0, \forall p$; $\sum_{j=1}^{J} b_{jq} > 0, \forall q$; and $\sum_{k=1}^{K} c_{kr} > 0, \forall r$).

Within the context of the ALS estimation of the three-mode partitioning model, one therefore has to detect the occurrence of empty clusters and subsequently undertake action to change this problematic situation somehow into one where empty clusters are no longer present.

In the present paper, three types of methods to deal with empty clusters are considered. In the following, one or more empty clusters are assumed to be present.

A first way for dealing with empty clusters is to search for the object that lies furthest away from its cluster centroid and to place this element in a singleton cluster. Furthest away is in this case formalized as follows: Consider the partition matrix **A**. If this matrix has a zero column, the element i of **A** is identified for which it holds that $\sum_{j=1}^{J}\sum_{k=1}^{K}(d_{ijk} - m_{ijk})^2$ is the highest. The empty cluster is then changed into the singleton cluster containing only element i. If more than one empty clusters were present, the element which lies second furthest away from its cluster centroid is put into the second empty cluster, and so on, until no further empty clusters are present. We will further refer to this procedure for handling empty clusters as the singleton procedure.

Kiers (2004) proposed another procedure for dealing with empty clusters in the three-mode partitioning problem. Consider the estimation of **A**, after each element is assigned to its closest cluster centroid, and one or more empty clusters are present, then it follows that the elements of the core $\underline{\mathbf{W}}$ that correspond to the empty clusters in $\mathbf{A}$ have no contribution to the loss function (4). They can thus be replaced by any value without affecting the loss function (4). Therefore, the author proposed to multiply these elements by -1, whereafter the partition matrix $\mathbf{A}$ is reestimated using the transformed core $\underline{\mathbf{W'}}$. If this procedure again yields empty clusters in $\mathbf{A}$, both $\mathbf{A}$ and the core are reset to their previous values, and thus no update of $\mathbf{A}$ has occurred. We will further refer to this procedure as the mirror procedure.

In the third approach to deal with empty clusters, Rocci and Vichi (2004) proposed a splitting procedure. Consider again the estimation of $\mathbf{A}$, if an empty cluster A_e turns up, the procedure used here identifies the elements i' that belong to the cluster A_m, the cluster that has the largest within cluster variability, and redistributes these elements by splitting them in two groups by applying a one-mode k-means clustering with $k = 2$ to the reduced data array containing only the elements i' identified earlier, and using an initial start for the latter procedure by assigning the first half of the set of elements i' to the first cluster, and the second half to the second cluster. If there are more than one empty clusters, this procedure is repeated until no more empty clusters are present. We will further refer to this procedure as the splitting procedure.

In general, heuristics for estimating a three-mode partitioning can be made up by a combination of any type of initialization, ALS-scheme and procedure to deal with empty clusters. One exception however occurs when it is chosen to use ALS-scheme $\mathbf{A}_{\underline{\mathbf{W}}}\mathbf{B}_{\underline{\mathbf{W}}}\mathbf{C}_{\underline{\mathbf{W}}}$. In this case no specific constraint needs to be imposed for preventing empty clusters because of the simple fact that due to continuous updating of the core this will not occur. Consider the situation where a cluster consists of one element only. When evaluating every possible allocation of this element (and allowing the core to drift correspondingly), it is always better, in terms of minimizing the loss function, to leave this element in the singleton cluster because in that case, the cluster mean will equal the observations, adding nothing to the loss function.

4 Simulation Study

Because the heuristic algorithms considered in this paper do not perform a full enumerative search on the entire solution space, we never know if a certain solution is truly the global optimum of the solution space. Therefore, in this section, we present a simulation study to examine the local minima problem for the three-mode partitioning model, when estimated by a number of ALS-algorithms put together using the elements of the framework presented in Section 3, under a number of conditions pertaining to the characteristics of the data. Second, we are also interested in the extent to which the optimal solutions resemble the true clustering structures underlying the data.

In Subsection 4.1, the design of the simulation study and the specific questions it is intended to answer are outlined. Next, the results are presented in Subsection 4.2 (local optima) and Subsection 4.3 (goodness of recovery).

4.1 Questions, Design and Procedure

The specific questions which we will try to answer with this simulation study are:

1. What is the performance, in terms of minimizing the loss function (4), of several ALS-algorithms? Is there an effect of the size of the data set, the underlying true rank, the error level and/or the number of clusters?

2. To which extent does, for each given data set, the optimal solution (across all combinations of *Algorithm* and *Start*) resemble the true structure underlying the data?

In order to deal with these different questions, we set up a simulation study. To explain its design, three different types of real-valued $I \times J \times K$ arrays must be distinguished in this simulation study: a true array $\underline{\mathbf{T}}$, which can be represented by a three-mode partitioning model of rank r; a data array $\underline{\mathbf{D}}$, which is $\underline{\mathbf{T}}$ perturbed with error; and the model array $\underline{\mathbf{M}}$ yielded by an estimation ALS algorithm, which can be represented by a three-mode partitioning model of the same rank as the true array $\underline{\mathbf{T}}$.

The design of the simulation study was fully crossed, the factors being as follows:

(1) the *Size*, $I \times J \times K$, of $\underline{\mathbf{T}}$, $\underline{\mathbf{D}}$, and $\underline{\mathbf{M}}$, at 3 levels: $48 \times 48 \times 48$, $96 \times 48 \times 48$, $96 \times 96 \times 48$;

(2) the *Rank*, r, of the three-mode partitioning model for $\underline{\mathbf{T}}$, at 5 levels: $2 \times 2 \times 2$, $2 \times 2 \times 4$, $4 \times 2 \times 2$, $2 \times 4 \times 4$, $4 \times 4 \times 4$;

(3) the *Error*, ε, which is the proportion of variance in $\underline{\mathbf{D}}$ unaccounted for by $\underline{\mathbf{T}}$, or the proportion of random noise present in the data, at 5 levels: .00, .01, .05, .20, .60;

(4) the *Equality of Cluster Sizes*, the number of modes that contain a cluster of considerable smaller cardinality than the other clusters pertaining to the same mode. At level one, all clusters corresponding to the same mode contain an equal amount of elements. At level two, the cardinality of one object cluster is five times as small as the cardinality of the other clusters of the objects mode while all cluster sizes corresponding to the other two modes are equal. At level three, both the elements of the objects and attributes mode are distributed amongst a number of clusters in which one has a cluster size five times as small as the other clusters. Finally, at level four, all three modes correspond to a set of clusters in which one is five times as small as the other clusters pertaining to the same mode.

For each combination of these four independent variables, 3 replicates were studied, yielding 3 × 3 (*Size*) × 5 (*Rank*) × 5(*Error*) × 4(*Equality of Cluster Sizes*)= 900 simulated data sets.

For each combination of the levels of *Size*, *Rank*, *Error* and *Equality of Cluster Sizes*, the 3 true arrays $\underline{\mathbf{T}}$ corresponding to the 3 replicates were constructed as follows: Partition matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are held constant within the same condition. The entries in the core array $\underline{\mathbf{W}}$ were all independent realizations of a uniformly distributed variable on the real interval [0,1]. The true array $\underline{\mathbf{T}}$ resulted from combining $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{W}}$ by (1). Subsequently, a data array $\underline{\mathbf{D}}$ was constructed by adding error to each true array $\underline{\mathbf{T}}$ using the following expression:

$$d_{ijk} \quad = \quad \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}w_{pqr} + e_{ijk} \qquad \forall i,j,k \tag{10}$$

where e_{ijk} is sampled from $N(0,\sigma^2)$. Note that if *Error* equals ϵ, then

$$\sigma \quad = \quad (\text{variance of the true array } \underline{\mathbf{T}} \times \frac{\epsilon}{1-\epsilon})^{\frac{1}{2}}$$

Finally, all data arrays $\underline{\mathbf{D}}$ were analyzed in the true rank by each combination of the following two factors:

(5) the *Type of Start*, at 4 levels: *Random Multi-Start*, *Independent k-means Multi-Start*, *Tucker3 Multi-Start* and *Sequential k-means Multi-Start*; For each data set and each of these four different types of initial starts, 50 multistarts were generated by the procedures as discussed in Subsection 3.1. Aside from these multistarts, we will also report on the analyses using the three types of single rational starts (*Independent k-means Start*, *Tucker3 Start* and *Sequential k-means Start*) discussed in Subsection 3.1.

(6) the *Algorithm*, at 7 levels: As depicted in Figure 3, a subset of 6 of these algorithms exist of taking all combinations of the first two ALS-schemes as described in Subsection 3.2 and the three types of procedures to deal with empty clusters as described in Subsection 3.3. *MIRR1* was proposed by Kiers (2004) and *SPLIT2* by Rocci and Vichi (2004). *DRIFT* was proposed by Schepers and Van Mechelen (2004) and uses the third ALS-scheme ($\mathbf{A}_{\underline{\mathbf{W}}}\mathbf{B}_{\underline{\mathbf{W}}}\mathbf{C}_{\underline{\mathbf{W}}}$...) as described in Subsection 3.2. For the estimation of any one of the individual cluster matrices simultaneously with $\underline{\mathbf{W}}$, the particular procedure in *DRIFT* takes on a similar form as the k-means algorithm described by Hartigan (1975) for the case of one-mode partitioning of a two-way data matrix. More in particular, consider the simultaneous estimation of $\mathbf{A}$

> and $\underline{\mathbf{W}}$. Starting with the first object in $\mathbf{A}$, every possible move of this object to another cluster, together with the corresponding update in $\underline{\mathbf{W}}$, is considered and the loss function (9) is evaluated. The move that decreases the loss function the most is taken as the new estimates of $\mathbf{A}$ and $\underline{\mathbf{W}}$ and the algorithm proceeds to the next object where the same process is repeated. After the algorithm has passed through all objects, two possible situations arise: either there has been a decrease in the loss function during the pass through all the objects and the procedure is repeated, or there has been no decrease in the loss function during the last pass through all the objects, the procedure stops and moves on to the next partition matrix. Note that these 'iterations within one mode' are possible because of the fact that any reallocation of any object of $\mathbf{A}$ can possibly affect the best assignment of any other object in $\mathbf{A}$ to a certain cluster. Note further that due to this dependency, the issue of choosing an initial estimate of $\mathbf{A}$ becomes relevant whenever the ALS-scheme moves from one partition matrix to another. In this algorithm, at iteration $n > 1$, the previous estimate $\mathbf{A}^{n-1}$ is taken as a starting configuration. The conditional estimation of $\mathbf{B}$ and $\mathbf{C}$ is performed completely analogous to that of $\mathbf{A}$. Finally, the algorithm stops when no further decrease in the loss function (4) is observed after the successive estimation of all three partition matrices.

Each analysis per data set yields 7 (*Algorithm*) × 4 (*Type of Start*) × 50 (*multistarts*) model arrays $\underline{\mathbf{M}}$ for each data array $\underline{\mathbf{D}}$. As a result of this procedure 3 (*replicates*) × 3 (*Size*) × 5 (*Rank*) × 5 (*Error*) × 4 (*Equality of Cluster Sizes*) × 7 (*ALS-scheme*) × 4 (*Type of Start*) × 50 (*multistarts*) different triplets ($\underline{\mathbf{T}}$,$\underline{\mathbf{D}}$,$\underline{\mathbf{M}}$) were obtained.

4.2 Minimization of the loss function

In this section, the local minima problem for three-mode partitioning is investigated and related to *Size*, *Rank*, *Error* and *Equality of Cluster Size* on the data side, and *Algorithm* and *Type of*

Start on the estimation side.

We discuss the results for analyses in true rank because we know that in that specific case, the error level is an upper bound for the *BOF* of the global optimum.

First, we examine for how many data sets it holds that at least one of the combinations of the levels of *Type of Start* and *Algorithm*, a solution was found that satisfies the criterion of having a *BOF* less than or equal to the error level in the data. It turns out that for all data sets, this was the case. As a consequence, we will further use the *BOF* of the best solution as a proxy for the *BOF* of the globally optimal solution for a given data set.

Second, we counted for each *Type of Start/Algorithm* combination and each data set how many times this optimal solution was found within the 50 multistarts. The higher this number, the more likely the optimal solution is found within any run of the multistart procedure, and thus the better the *Type of Start/Algorithm* combination is at minimizing the loss function (4).

An analysis of variance with the latter measure as dependent variable; *Size*, *Rank*, *Error* and *Equality of Cluster Sizes* as between-subjects variables and *Type of Start* and *Algorithm* as within-subjects variables revealed an intraclass coefficient ρ_I (Haggard, 1958) of .46 for the effect of *Algorithm.* Regarding only effect sizes larger than .10 as sizeable, *Rank* ($\rho_I = .28$) and the interaction between *Rank* and *Algorithm* ($\rho_I = .11$) remain. All other effects together only explain the remaining 16% of the variance in the dependent measure.

Figure 4 shows that algorithms *SPLIT1* and *SPLIT2* are most likely to return the optimal solution and that their performance is near perfect. Also, the optimal solution is less likely to be found when *Rank* increases, and this is more pronounced for algorithms *MIRR1* and *MIRR2*, whereas algorithms *SPLIT1* and *SPLIT2* are far less sensitive to this effect. These results suggest that it is best to use either one of algorithms *SPLIT1* or *SPLIT2.* The high performance of these two algorithms further suggests that one should suffice with some 5 multistarts (be it random or rational) to find the best solution, given a data set.

insert Figure 4 here

In order to investigate the quality of the single rational starts (*Independent k-means Start*, *Tucker3 Start* and *Sequential k-means Start*), we investigated, for each data set, whether or not the *BOF* of the solution returned using one of these types of starts is equal to the *BOF* of the optimal solution as identified earlier. Overall, for all the data sets analyzed by using *Independent k-means Start* as initialization, followed by an analysis using each *Algorithm*, 52% of these analyses ended in a solution as good as the optimal solution. For *Sequential k-means Start* and *Tucker3 Start*, this was equal to 73% and 99%, respectively. Apparently, using the single rational *Tucker3 Start* provides a very good initialization for estimating a three-mode partitioning, since it leads to the optimal solution in almost every case.

4.3 Goodness of Recovery

With respect to evaluating the agreement of the partitioning structures between the true underlying partition of the set of objects (resp. attributes, sources) in the three-mode partitioning model for $\underline{\mathbf{T}}$ and the estimated partition of the set of objects (resp. attributes, sources) for the optimal solution (over all combinations of *Type of Start* and *Algorithm*), for each data set, the corrected Rand index (Hubert & Arabie, 1985) was used. This index equals 1 if the two partitions are identical and 0 if the two partitions do not correspond more than expected by chance. A combined corrected Rand index (c-CRI) was calculated by taking the average corrected Rand index for the object, the attribute and the source partitions, weighted by the number of objects, attributes and sources, respectively.

Over all 900 data sets, in none of the cases the c-CRI between the true partitioning and the partitioning corresponding to the optimal solution was not equal to 1, implying that the true partitions and the estimated partitions of the best fitting model are identical in every case. Note that in this simulation study, the highest level of ϵ is .60, implying that the data are quite far from the truth. The results found here can not be attributed to an artificial effect due to manipulating an insufficient range of *Error*.

5 Four Empirical Data Sets

5.1 Minimization of the loss function

In this section we report the results of analyses of four empirical three-way three-mode data sets by each combination of *Algorithm* and *Type of Start*. These data sets were the following ones:

1. Anger-Consequence data (Van Coillie & Van Mechelen, 2004): A set of 139 respondents indicated for a set of 8 consequences to what extent they expected them to change after they would have executed each behavior out of a set of 16 anger-related behaviors.

2. Anger data (Kuppens & Van Mechelen, in press): Respondents indicated to what degree they experienced a list of 24 anger related responses in 14 different situations. Leaving out all subjects for which missing values are observed, 357 respondents remain.

3. Chopin data (Murakami & Kroonenberg, 2003): 38 Japanese university students rated 24 short piano solo pieces composed by Frederik Chopin on a set of 20 bipolar scales.

4. Archetypal patients (Mezzich & Solomon, 1980): Each of 22 psychiatrists was invited to think of a typical patient for each of 4 diagnostic categories and characterize each patient in terms of severity on 17 psychiatric symptoms.

All data sets were analyzed in each level of *Rank* by taking all possible combinations of 2 to 5 clusters for each mode except for the patient mode in the archetypal patient data set for which the highest number of clusters in the analyses was limited to 3 due to the small number of elements in this mode. For each rank, every combination of *Algorithm* and *Type of Start* was run, each using 50 multistarts out of which the best solution found was retained. Next, for each combination of

Algorithm and *Type of Start*, it was examined how many times the optimal solution was found within these 50 multistarts.

An analysis of variance performed per data set, with the levels of *Rank* as blocks and *Type of Start* and *Algorithm* as fully crossed independent variables, shows a significant effect of *Algorithm* ($p < .0001$) but not for *Type of Start*, for each data set. This confirms what we expected from the results of the simulation study. Figure 5 shows the box plots of the number of times the best solution was found by each algorithm, over all levels of *Rank* and *Type of Start*, for each data set. A surprising finding when analyzing any one of these four empirical data sets is the fact that algorithm *DRIFT* now does a better job than algorithms *SPLIT1* and *SPLIT2*. Moreover, the performance of any of the algorithms other than *DRIFT* is highly unsatisfactory.

5.2 Parametric bootstrapping

In the following, we present a further investigation by highlighting one of the data sets presented in Subsection 5.1, the Anger data (Kuppens & Van Mechelen, in press). We will focus on only one instance of *Rank*, $2 \times 2 \times 4$, because this is one of the levels of *Rank* where the reversal effect in performance of the algorithms is particularly clear: *DRIFT* returns the optimal solution in almost every run here, while the other algorithms almost never do. In order to get an idea of what might cause this reversal of performance as compared to the results of the simulation study, we investigated in what way the optimal solution corresponding to this data set differs from the ones in the simulation study.

Recall that the stochastic version of the model with independently and identically distributed normal error distributions for each data cluster leads to the same loss function as the deterministic model (4). The artificial data sets described in the simulation study in Subsection 4.1 were generated in accordance to this stochastic model formulation. In order to investigate the influence of departures from the assumptions in the stochastic model, we assessed to what extent the different subpopulations corresponding to the data clusters returned in the optimal solution of rank $2 \times 2 \times 4$ of the Anger data differed from the stochastic model assumptions implied by (6). More in particular, we investigated (i) the within-cluster variance, (ii) the within-cluster skewness, and (iii) the proneness to outliers of the within-cluster residual distributions. Finally, we explored (iv) the covariance structure between the residuals.

To examine possible discrepancies between the optimal solution given the observed data set and the optimal solution under the null model with iid $N(0, \sigma^2)$ distributed residuals, we used a parametric bootstrapping procedure. First, we obtained the reconstructed data $\underline{\mathbf{M}}$ by using the estimates $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{W}}$, returned by the optimal solution. Then, we estimated the pooled within-cluster variance corresponding to the optimal solution, $\hat{\sigma_\epsilon}$, and drew 1000 bootstrap residual distributions with all values iid from $N(0, \hat{\sigma_\epsilon})$. By adding these bootstrap residual populations to $\underline{\mathbf{M}}$, we obtain 1000 bootstrap data sets that are all generated from the null model.

(i) with respect to the variance of the residuals, we wanted to investigate whether the within-cluster variances were more variable in the empirical data set as compared to data generated under the null model. We computed the ratio of the highest within-cluster variance over the smallest within-cluster variance for the empirical data set and for all bootstrap data sets. (ii) With respect to skewness, we calculated the average skewness of the distributions of residuals corresponding to each data cluster, for the empirical data set and for all bootstrap data sets. (iii) With respect to proneness to outliers, we calculated the average kurtosis value over the distributions of residuals corresponding to each data cluster. Finally, (iv) we calculated the variance-covariance matrix of the $I \times JK$ matricized array of residuals, and summed the squared values of the above-diagonal elements of this matrix. Figure 6 shows the results for these four statistics calculated for the bootstrap data sets (boxplots), and the empirical data set (stars).

insert Figure 6 here

Clearly, the values of all 4 statistics calculated for the Anger data are highly unlikely to be observed should the residuals in fact be drawn in accordance to the stochastic model assumptions. We conclude that the within-cluster variances corresponding to the Anger data are not equal, that the within-clusters distributions of residuals are on average skewed and more prone to outliers, and finally that there is a covariance between the residuals stemming from elements of the third and second mode. These types of violations are generally observed in analyzing the four empirical data sets and it rises the suspicion that one of these violations, or a combination of several of them, may affect algorithmic performance. In order to get more insight into this, a second simulation study was conducted which is discussed in the next section.

6 Second Simulation Study

In order to test whether violations of the implicit stochastic model could affect algorithmic performance, we generated a new group of artificial data sets, each of them in some way violating the assumptions of model (6).

All data sets had *Size* $= 20\times20\times20$, *Rank* $= 4\times4\times4$, and *Error* $= 0.3$. Table 1 shows the list of distributions from which the error was drawn, and the value of their respective parameters. After generating the residuals from these distributions, they were rescaled in order to control for the error level in the data. Within each level of *Residual Distribution*, three different levels of *Inequality in within cluster variance*, pertaining to the number of clusters with different within-cluster variances, were generated. Either 0, 16 or 32 out of the total of 64 clusters had a within-cluster variance which was 9 times as large as the within-cluster variance of the remaining clusters. For each combination of *Residual Distribution* and *Inequality in within cluster variance*, 10 data sets were generated. Each data set was then analyzed with 50 random multistarts by each of the seven algorithms. We only used random starts in order to reduce computational costs and because we could not find an effect of *Type of Start* in the previous investigations.

insert Table 1 here

In all of the above conditions, the same pattern of performance as in the first simulation study was found. This suggests that the ordering of algorithms in terms of performance in minimizing the loss function is robust against violations of the implicit stochastic model, such as different within-cluster variances, more outlier prone and more skewed residual distributions.

Finally, an additional condition in which the independence assumption was violated, was examined. For the data sets in this last condition, the residuals were drawn from the normal distribution as in (10), but now with a nondiagonal covariance matrix (in particular, a variance-covariance matrix of all attribute-source combinations with anti-robinson form, with successive entries in this matrix, when moving away from the main diagonal, defined by linearly spaced intervals between 1 and 0). Figure 7 shows the results on 100 data sets analyzed in this condition. It can be seen that *DRIFT* returns the optimal solution more frequently than the other algorithms. Particularly striking is also the large decrease in performance (as compared to the results of the first simulation study) of algorithms *SPLIT1* and *SPLIT2*. As is the case for the four empirical data sets, a reversal of the ordering of the different algorithms in terms of performance in minimizing the loss function can be observed.

insert Figure 7 here

7 Conclusion and Discussion

This paper recapitulates the three-mode partitioning model, which yields a simultaneous partitioning of all three modes of a three-way three-mode data array. This paper further presented a framework for alternating least-squares algorithms for fitting a three-mode partitioning model to a data set.

A simulation study was conducted to examine whether the model can be estimated and also to evaluate and compare seven different algorithms, combined with four different initializations. For the results of the first simulation study, a comparison between the seven algorithms with respect to their *BOF* showed that algorithms *SPLIT1* and *SPLIT2* were most successful in minimizing the loss function since they succeeded in finding the optimal solution in almost every run of a multistart procedure. Generally, the results discussed in this paper always point to the fact that it does not matter whether the core is updated after each estimation of a partition matrix or after the successive estimation of all three partition matrices. What does matter is the procedure for dealing with empty clusters. Apparently, the splitting procedure for dealing with empty clusters is very efficient with respect to finding the optimal solution of the solution space. The mirror procedure and to a lesser extent the singleton procedure provide worse results.

Next to *Algorithm*, only *Rank* and the interaction between *Rank* and *Algorithm* showed up as important effects in explaining the variability in the number of times the best solution was found. Figure 4 showed that this is due to the fact that all algorithms, but to a lesser extent algorithms *SPLIT1* and *SPLIT2*, perform worse when the underlying rank of the data increases. All other factors such as *Type of Start*, *Size*, *Error* and *Equality of Cluster Sizes* did not turn out to be important factors in determining the number of times the best solution was found within a multistart procedure.

The results of analyses on a number of real data sets provide a different perspective. For these data sets, it turns out that algorithm *DRIFT* returns a better solution more frequently than algorithms *SPLIT1* and *SPLIT2*. A second simulation study in which all conditions in some way violated the assumptions of the stochastic equivalent of (4), showed that in the case were additional systematic error in the form of a covariance was present, we were able to replicate a similar reversal in performance of the algorithms as observed in the real data sets. Considering the fact that often three-way three-mode data sets are gathered through questionnaires in which respondents sequentially answer a set of questions, some form of additional dependency, such as a covariance between successive questions, can be expected quite frequently. Therefore, these results suggest that in a practical situation, where we do not know the characteristics of the individual cluster distributions in advance, it is best to use a combination of both *DRIFT* on the one hand and *SPLIT1* or *SPLIT2* on the other hand, when given an empirical data set. Note that in using the first, one does have to accept a larger computation time since this algorithm generally takes more time to converge to a stable solution.

Possibly, the violation on the stochastic model as described in the covariance condition and as observed in a number of real empirical data sets, in some way generates a specific structure in the solution space that causes the algorithms that no not continuously update the core to end up more often in locally suboptimal solutions whereas algorithm *DRIFT* does not seem to suffer from the same problem. We are not yet able to explain this finding.

Finally, the analyses performed on four empirical data sets also indicate that one should not be too economical with respect to the amount of multistarts. In contrast to the results of the simulation study, experience on four empirical data sets discussed in this paper suggest that at least 50 multistarts is not always an exaggerated amount.

As a possible direction of future research, performance of other types of optimization algorithms such as genetic algorithms, simulated annealing and other local search strategies, for an overview see Aarts and Lenstra (1997), might be explored and compared to the algorithms discussed in this paper. We would like to emphasize that the latter should be done not only on artificial data but also on a range of real data sets since our experience suggests that quite different pictures might evolve from both approaches.

References

Aarts, E., & Lenstra, J. (1997). *Local search in combinatorial optimization.* New York: Wiley.

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* (pp. 557–566). Berlin, Germany: Springer.

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, *49*, 803-821.

Bock, H. H. (1996). Probability models and hypotheses testing in partitioning cluster analysis. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 377–453). River Edge, NJ: World scientific Publ.

Bryant, P., & Williamson, J. A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika*, *65*, 273-281.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology*, *31*, 607–649.

Celeux, G., & Govaert, G. (1992). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and simulation*, *14*, 315-332.

Chaturvedi, A., & Carroll, J. D. (1994). An alternating combinatorial optimization approach to fitting the INDCLUS and generalized INDCLUS models. *Journal of Classification*, *11*, 155–170.

De Boeck, P., & Rosenberg, S. (1988). Hierarchical classes: Model and data analysis. *Psychometrika*, *53*, 361–381.

Gaul, W., & Schader, M. (1996). A new algorithm for two-mode clustering. In H. Bock &

W. Polasek (Eds.), *Classification and knowledge organization* (pp. 15–23). Berlin, Germany: Springer.

Govaert, G. (1995). Simultaneous clustering of rows and columns. *Control and Cybernetics*, *24*, 437-458.

Govaert, G., & Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, *36*, 463-473.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance.* New York: Dryden.

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

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

John, S. (1970). On identifying the population of origin of each observation in a mixture of observations from two normal populations. *Technometrics*, *12*, 553-563.

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., & De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, *45*, 69–97.

Kuppens, P., & Van Mechelen, I. (in press). Determinants of the anger appraisals of threatened self-esteem, other-blame, and frustration. *Cognition & Emotion.*

McLachlan, G. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. In P. R. Krishnaiah & L. N. Kanal (Eds.), *Handbook of statistics (vol.2)* (pp. 199–208). Amsterdam: North-Holland.

Mezzich, J. E., & Solomon, H. (1980). *Taxonomy and behavioral science: comparative performance of grouping methods.* London: Academic Press.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*, 325-342.

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

Murtagh, F. (1989). (review of the book *Data, Expert Knowledge and Decisions*). *Journal of Classification*, *6*, 129-132.

Rocci, R., & Vichi, M. (2003). *Three-mode clustering of a three-way data set.* (Paper presented at CLADAG 2003, Bologna)

Rocci, R., & Vichi, M. (2004). *Multimode partitioning.* (Submitted for publication)

Schepers, J., & Van Mechelen, I. (2004). *Three-mode partitioning: Model and algorithm.* (Paper presented at gfkl 2004, Dortmund)

Scott, A. J., & Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, *27*, 387–397.

Selim, S., & Ismail, M. (1984). K-means type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 81-87.

Steinley, D. (2003). Local optima in k-means clustering: What you don't know may hurt you. *Psychological Methods*, *8*, 294-304.

Van Coillie, H., & Van Mechelen, I. (2004). *Expected consequences of anger-related behaviors.* (Submitted for publication)

Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research*, *13*, 363-394.

Vichi, M. (2002). 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* (pp. 43–51). Heidelberg, Germany: Springer.

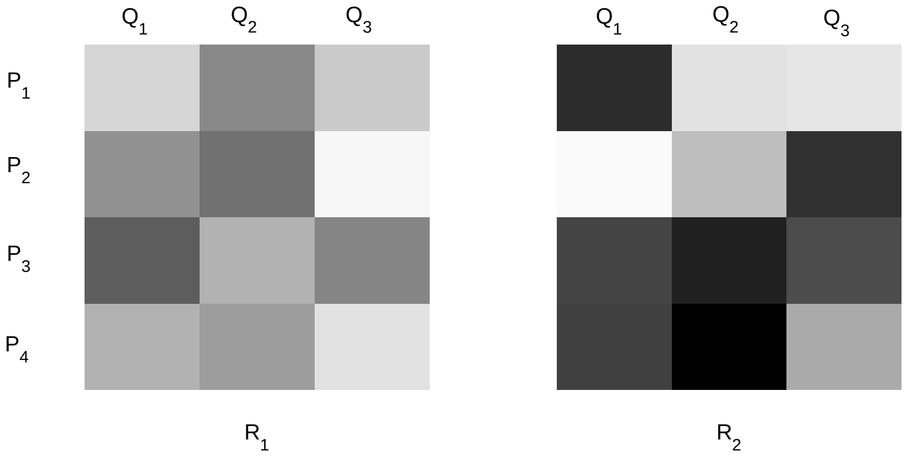

Figure 1: heat map of a hypothetical three-mode partitioning with nr of object clusters = 4, nr of attribute clusters = 3 and nr of source clusters = 2.

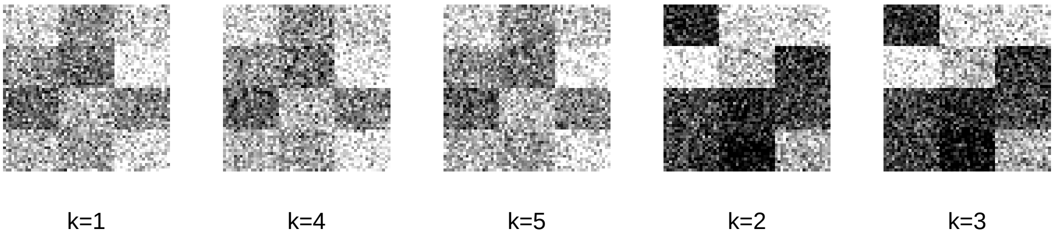

Figure 2: Non-destructive heatmap of the same hypothetical example as in figure 1. In this case, the number of sources (cardinality of the third mode) is equal to 5.

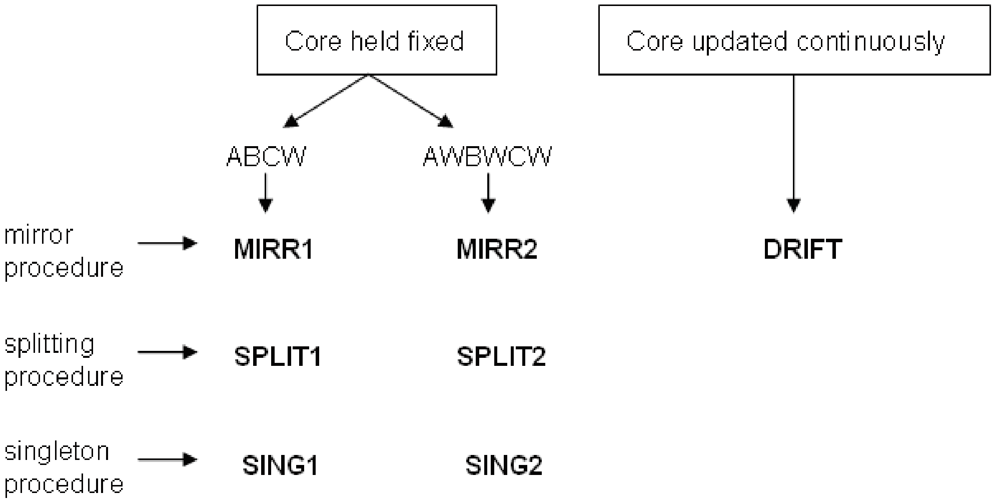

Figure 3: Graphical representation of how the seven algorithms are situated within the ALS framework.

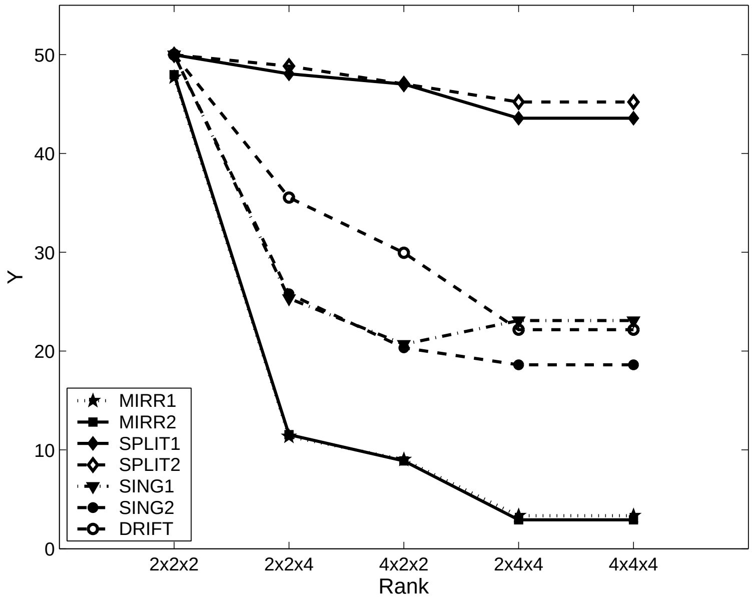

Figure 4: Line plots of the mean number of times (Y) the optimal solution is found within 50 runs, for each *Algorithm* and each level of *Rank*.

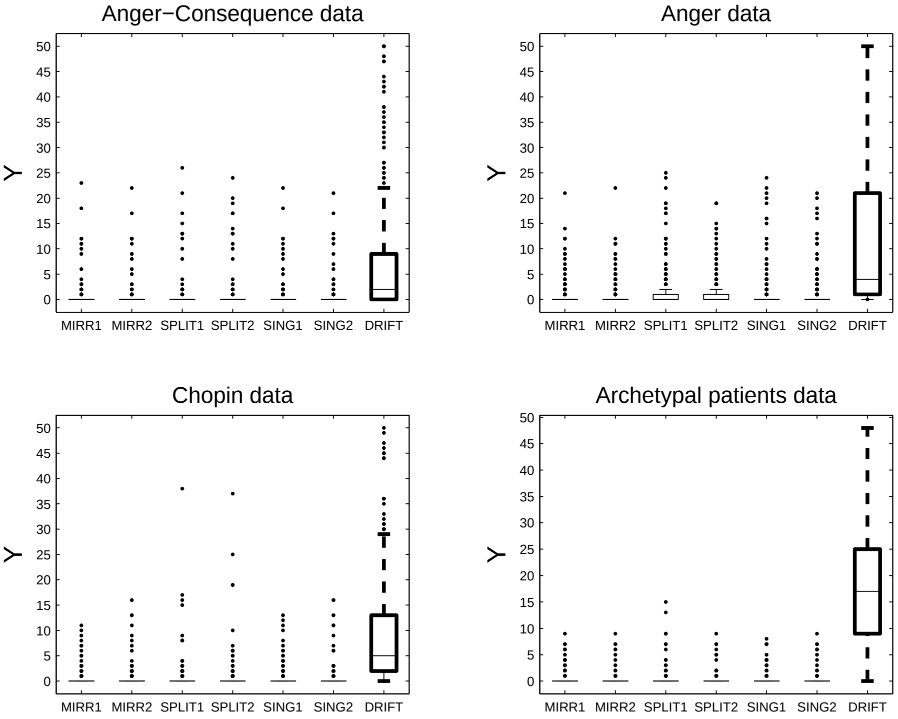

Figure 5: Box plots of the number of times (Y) the best solution was found, over all combinations of ranks, by each of the different algorithms, for each empirical data set.

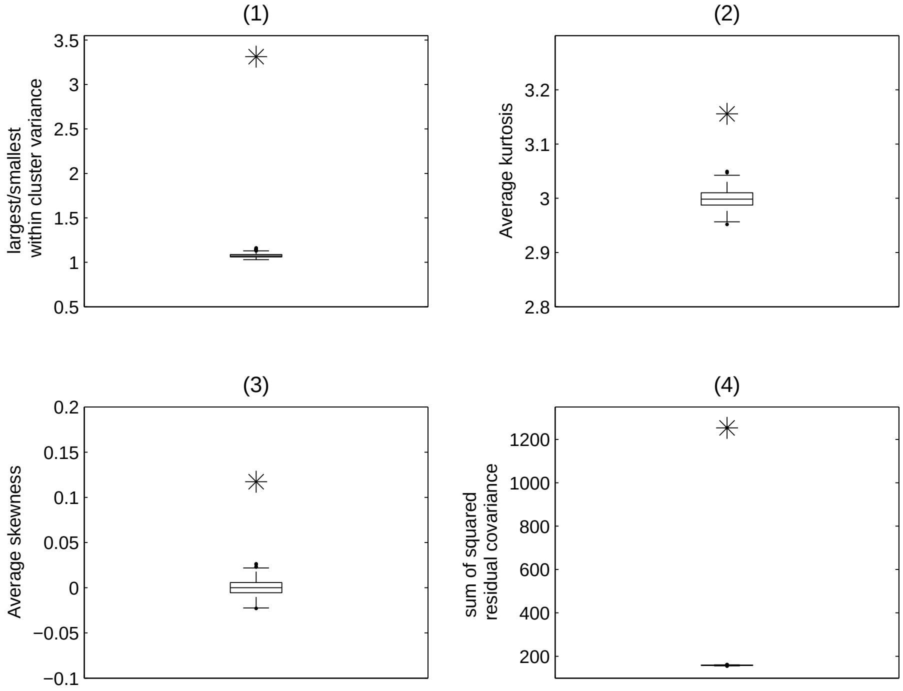

Figure 6: Box plots of the parametric bootstrap tests for (1) the ratio of the largest versus the smallest within-cluster variances, (2) the average within-cluster kurtosis of the residual distributions, (3) the average within-cluster skewness of the residual distributions, and (4) the sum of squared residual covariances. The $*$ indicates the value of the calculated test statistic for the Anger data.

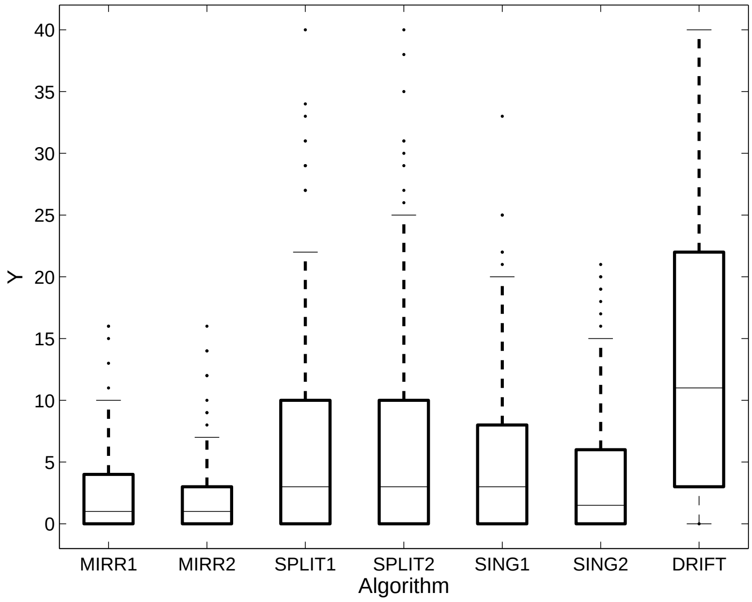

Figure 7: Box plots of the number of times (Y) the best solution was found by each of the different algorithms in the covariance condition.

Table 1: distributions used in the second simulation study

distribution	parameters	
t	$\nu = 5$	symmetric, thicker tails than normal distribution
beta	$\alpha = 2, \beta = 3$	skewed to the right
lognormal	$\mu = .5, \sigma = .3$	skewed to the right
poisson	$\lambda = 2$	skewed to the right, non-continuous
uniform	$\alpha = -.25,\ \beta = .25$	symmetric
weibull	$\beta = 1,\ \gamma = 2$	skewed to the right
χ^2	$\nu = 4$	skewed to the right