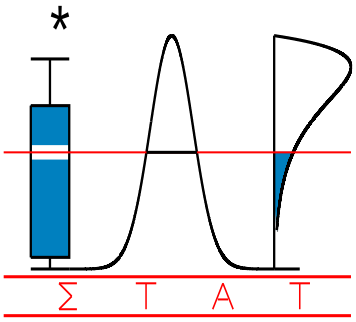
T E C H N I C A L
R E P O R T

0352

K-CENTROIDS HIERARCHICAL CLASSES ANALYSIS

LOMBARDI, L., CEULEMANS, E. and I. VAN MECHELEN



I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

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

K-centroids Hierarchical Classes Analysis

Luigi Lombardi

Department of Cognitive Science and Education, University of Trento,

via Matteo del Ben, 5, I-38068 Rovereto (TN), Italy. email: lombardi@form.unitn.it

Eva Ceulemans

Department of Psychology, Catholic University of Leuven,

Tiensestraat 102, B-3000 Leuven, Belgium. email: eva.ceulemans@psy.kuleuven.ac.be

Iven Van Mechelen

Department of Psychology, Catholic University of Leuven,

Tiensestraat 102, B-3000 Leuven, Belgium. email: iven.vanmechelen@psy.kuleuven.ac.be

November 10, 2003

Author Notes: This work was done while the first author was international scholar at the Catholic University of Leuven. Correspondence concerning this paper should be addressed to Luigi Lombardi, Dipartimento di Scienze della Cognizione e della Formazione, via Matteo del Ben, 5, I-38068 Rovereto (TN), Italy. Email: lombardi@form.unitn.it.

Abstract

In this paper we present a new model for binary two-way two-mode clustering, called K-centroids hierarchical classes model (KC-HICLAS). KC-HICLAS is a new member of the family of hierarchical classes models (HICLAS) (De Boeck and Rosenberg, 1988). Like any HICLAS model, the KC-HICLAS model includes a hierarchical classification of the elements of each mode and a linking structure between the two hierarchies. Unlike other HICLAS models, KC-HICLAS explicitly limits the classification of the elements of one mode to maximum K (not necessarily distinct) classes. An algorithm to fit the KC-HICLAS model is described and evaluated in a simulation study. The model is then illustrated with two applications (one exploratory and one confirmatory) to psychiatric data sets. Finally, the relationships between KC-HICLAS and other clustering techniques, especially, K-means (MacQueen, 1967) and factorial K-means (Vichi and Kiers, 2001) are discussed.

Keywords: Binary data; Two-way Two-mode clustering; HICLAS models.

1 Introduction

Various techniques for two simultaneous and linked classifications have been developed within the context of two-mode clustering (for overviews, see, e.g., Eckes, 1991; Eckes and Orlik, 1993; Van Mechelen, Bock and De Boeck, 2003). These techniques are often of substantive interest for both biological and social sciences and several applications in biometrics, psychometrics and sociometrics have been documented in the literature (Baier et al., 1997; Gaul and Schader, 1996; Getz et al., 2000). Within the two-mode clustering family, several methods seek for a partition of each of the modes involved in the data. In the latter case, for example, in a patient by psychiatric symptom data matrix each patient is assigned to a particular diagnostic category, whereas each symptom is assigned to a specific symptom family.

Most standard two-mode partitioning methods do not provide any structural organization of the sought partitions. Yet, researchers frequently need to recover (1) structural information within the mode–partitions and (2) linking relations between the partitions of the different modes. As such, in the above example, for diagnostic purposes, one may wish to order the diagnostic categories with respect to their linked symptom clusters; in that case, a diagnostic category would show up as more severe than another whenever the set of symptoms associated with the first is a superset of the set of symptoms associated with the latter.

HICLAS models (De Boeck and Rosenberg, 1988; Van Mechelen, De Boeck, and Rosenberg, 1995; Leenen, Van Mechelen, De Boeck, and Rosenberg, 1999; Ceulemans, Van Mechelen, and Leenen, in press; Ceulemans and Van Mechelen, in press) are structural models for N–way N–mode data that imply linked partitions (as well as linked overlapping clusterings) of the N modes under study. Unlike in many other methods, the partition classes of a given mode as implied by HICLAS are structurally organized in terms of if–then type partial orders. Up to now, the hierarchical classes family has been limited to models that imply only an implicit upper bound on the cardinalities of the partitions of the distinct modes in terms of a function of the so–called

mode–ranks of the models. Yet, researchers sometimes may wish to limit explicitly the classification of the elements of one of the modes to at most K partition classes (whereas the partitioning of the other modes is left unconstrained). Two substantive reasons may call for the latter scenario: Firstly, for descriptive purposes, a researcher may wish to limit the size of the partition of one mode in order to obtain a simpler hierarchical structure for that mode. Alternatively, in a confirmatory approach, a researcher may have an a priori model of the partitioning schema of the selected mode in terms of the number of classes in the partition; in this second case, the main goal is to test the goodness of fit of the constrained model with respect to the data.

In this paper we present a new model for clustering of two-way two-mode binary data. The model, which we call KC-HICLAS (K-Centroids Hierarchical Classes Model), is a novel extension of the hierarchical classes model family. Similar to all the members of this family it includes linked structurally organized classifications of the two modes. However, unlike the existing HICLAS models, KC-HICLAS explicitly limits the partition of the elements of one mode to maximum K (not necessarily distinct) classes.

In order to provide a self–contained exposition, the next section (Section 2) briefly recapitulates the main aspects of the standard hierarchical classes model. In Section 3 and Section 4 we present the new model and the associated algorithm, respectively. In Section 5, the results of a simulation study to evaluate the algorithm's performance are reported. Section 6 presents two applications to real data sets. Finally, Section 7 discusses the relation of the new model with extant partitioning models and introduces some possible useful extensions.

2 Standard HICLAS Model

In this section we take the original disjunctive hierarchical classes (HICLAS) model as defined by De Boeck and Rosenberg (1988) as a starting point.

A HICLAS analysis approximates an I (object) $\times J$ (attribute) binary data matrix $\mathbf{D}$ by an $I \times J$ binary reconstructed data or model matrix $\mathbf{M}$ that can be decomposed into an $I \times R$ binary matrix $\mathbf{A}$ and an $J \times R$ binary matrix $\mathbf{B}$, where R denotes the rank of the model. $\mathbf{A}$ includes R binary column vectors, called object bundles, and hence is called the object bundle matrix. Similarly, $\mathbf{B}$ includes R binary attribute vectors, called attribute bundles, and hence is called the attribute bundle matrix. Three types of relations among the rows (objects) and columns (attributes) of $\mathbf{M}$ are represented by the bundle matrices $\mathbf{A}$ and $\mathbf{B}$: association, equivalence, and hierarchy. As a guiding example we use the hypothetical matrices shown in Table 1-3.

2.1 Association relation

The association relation in $\mathbf{M}$ is the binary relation between the object mode and the attribute mode of $\mathbf{M}$ as defined by the 1-entries of $\mathbf{M}$. By the representation of this relation, $\mathbf{M}$ may be fully reconstructed from the hierarchical classes model. More precisely, in the case of the disjunctive model we have that

$$\mathbf{M} = \mathbf{A} \otimes \mathbf{B}^{'} \tag{1}$$

where $^{'}$ and $\otimes$ denote transpose and the Boolean matrix product (Kim, 1982), respectively. This association rule means that for an arbitrary entry m_{ij} of $\mathbf{M}$,

$$m_{ij} = 1 \Leftrightarrow \exists r, \quad 1 \leq r \leq R \quad : \quad a_{ir} = 1 \quad \text{and} \quad b_{jr} = 1. \tag{2}$$

For example, from the model in Table 3, it can be derived that Object 3 is associated with Attribute e, because both elements belong to Bundle III.

2.2 Equivalence relations

Equivalence relations are defined among the elements of each mode. Objects are equivalent iff they are associated with the same set of attributes. Likewise, attributes are equivalent iff they are

associated with the same set of objects. Sets of equivalent objects (attributes) are called object (attribute) classes. In the disjunctive HICLAS model, equivalent elements have identical bundle patterns. In the example in Table 2, Object 3 and 4 are associated with the same set of attributes, and, hence, those objects are equivalent and have identical bundle patterns in the HICLAS model of Table 3. At the attribute side, c and d are equivalent elements. Clearly, the object (attribute) classes constitute a partitioning of the object (attribute) mode. As a consequence, because of the representation of the equivalence relations, HICLAS may be considered a two-mode partitioning technique. Note in this regard that the HICLAS model rank R implies an upper–bound to the number of object (resp. attribute) partition classes. More precisely, the number of classes within the object (resp. attribute) partition of an HICLAS model ($\#C$) cannot exceed the cardinality of the hypercube $\{0,1\}^R$, that is,

$$\#C \leq 2^R. \tag{3}$$

2.3 Hierarchical relations

Hierarchical relations are defined among the elements (or classes) of each mode. An object is hierarchically below another, iff the respective sets of associated attributes are in a subset/superset relation. Similarly, an attribute is hierarchically below another, iff the respective sets of associated objects are in a subset/superset relation. In the disjunctive HICLAS model, hierarchical relations are represented directly in the corresponding bundle matrix in that an object is hierarchically below another iff the bundle pattern of the first is a subset of the bundle pattern of the latter. The same reasoning can be applied at the attribute side. For example, in Table 2, Attribute b is hierarchically below Attribute c; as a consequence, the bundle patterns of attributes b and c are in a subset-superset relation in the HICLAS model of Table 3.

2.4 Graphic representation

De Boeck and Rosenberg (1988) describe a graphic representation of the disjunctive model that gives a comprehensive account of the model. Figure 1 presents a graphic representation of the disjunctive model of Table 3. The upper half of the figure represents the hierarchy of the row (object) classes and the lower half the hierarchy of the column (attribute) classes, the latter being represented upside down. The zigzags denote the bundles, all rows (columns) above (below) a zigzag constituting a row (column) bundle. The association relation can be read from the representation as follows: An object (attribute) is associated with all attributes (objects) to which it is linked via a downward (upward) path of links and zigzags.

3 The new KC-HICLAS model

Assume that we have an $I \times J$ binary data matrix $\mathbf{D}$, and let K be the number of partition classes being sought for the I elements of the first mode. Then a KC-HICLAS analysis approximates $\mathbf{D}$ by an $I \times J$ binary model matrix $\mathbf{M}$ that can be decomposed into an $I \times K$ binary matrix $\mathbf{P}$ and a $K \times J$ binary matrix $\mathbf{M}^*$. In particular, $\mathbf{P}$ is an indicator matrix for membership of the I objects in K mutually exclusive, non–overlapping clusters (so that each row of $\mathbf{P}$ has exactly one element equal to one, the remaining elements being equal to zero), and $\mathbf{M}^*$ is a matrix composed of K partition class centroids. Note in this regard that the centroid of a set of Boolean vectors is to be understood here as the Boolean vector that is as close as possible to the set in question in the least absolute deviations (or, equivalently, in the least squares) sense. Moreover, we also assume that $\mathbf{M}^*$ can be further decomposed into a $K \times R$ binary matrix $\mathbf{A}^*$ and a $J \times R$ binary matrix $\mathbf{B}^*$, where (R, K) denotes the rank of the model. The latter means that the partition class centroids as included in $\mathbf{M}^*$ are constrained to be located in the Boolean vector space spanned by the columns of $\mathbf{B}^*$. It should be noted that only the data matrix $\mathbf{D}$ is known, whereas $\mathbf{P}$, $\mathbf{A}^*$, and $\mathbf{B}^*$ must be

estimated. As a guiding example we use the hypothetical matrices shown in Table 4–6.

The matrices $\mathbf{P}$, $\mathbf{A}^*$ and $\mathbf{B}^*$ of a KC-HICLAS model represent three types of structural relations.

3.1 Association relation

The association relation in KC-HICLAS is represented as follows:

$$\mathbf{M} = \mathbf{P}(\mathbf{A}^* \otimes \mathbf{B}^{*\prime}). \tag{4}$$

This association rule means that for an arbitrary entry m_{ij} of $\mathbf{M}$,

$$m_{ij} = \sum_{k=1}^{K} p_{ik} \bigoplus_{r=1}^{R} a^*_{kr} b^*_{jr}, \tag{5}$$

where $\oplus$ denotes the Boolean sum (Kim, 1982). Equation (5) implies that for an arbitrary entry m_{ij} of $\mathbf{M}$,

$$\begin{aligned} m_{ij} = 1 \quad \Leftrightarrow \quad & \exists k, 1 \le k \le K; \exists r, 1 \le r \le R: \\ & p_{ik} = 1 \quad \text{and} \quad a^*_{kr} = 1 \quad \text{and} \quad b^*_{jr} = 1 \end{aligned} \tag{6}$$

For example, from the decomposition model in Table 5, it can be derived that Centroid α is associated with Attribute c, and therefore that all objects in Class $\{1, 2, 3\}$ are also associated with Attribute c.

3.2 Equivalence relations

In KC-HICLAS we consider equivalence relations among the centroids and among the attributes. As in standard HICLAS, equivalent centroids (resp. attributes) should have identical bundle patterns. More precisely, a centroid k (resp. an attribute j) is equivalent to another centroid k' (resp. another attribute j') if and only if $\mathbf{A}^*_{k:} = \mathbf{A}^*_{k':}$ (resp. $\mathbf{B}^*_{j:} = \mathbf{B}^*_{j':}$). For example, centroids α, β and γ are associated with different sets of attributes; hence, those centroids are not equivalent and have different bundle patterns in the KC-HICLAS model of Table 6.

Note that if two centroids k and k' are equivalent, then all objects belonging to the corresponding partition classes are equivalent as well. The latter type of equivalence may be read from the matrix $\mathbf{A} = \mathbf{P}\mathbf{A}^*$ in that if $\mathbf{A}^*_{k:} = \mathbf{A}^*_{k':}$ then $\mathbf{A}_{i_1:} = \mathbf{A}_{i_2:} (= \mathbf{A}^*_{k:} = \mathbf{A}^*_{k':})$ for all i with $p_{ik} = 1$ or $p_{ik'} = 1$.

3.3 Hierarchical relations

We further consider hierarchical relations among the centroids and among the attributes. In particular, a centroid k (resp. an attribute j) is hierarchically below another centroid k' (resp. another attribute j') if and only if $\mathbf{A}^*_{k:} \leq \mathbf{A}^*_{k':}$ (resp. $\mathbf{B}^*_{j:} \leq \mathbf{B}^*_{j':}$). For example, as appears from Table 5, Centroid β is hierarchically below Centroid α; consequently, the bundle patterns of Centroids β and α in the KC-HICLAS model of Table 6 are in a subset/superset relation. Moreover, like for the equivalence relations, if a centroid k is hierarchically below another centroid k' then all objects belonging to the cluster associated with the first are hierarchically below all objects belonging to the cluster associated with the latter.

4 KC-HICLAS data analysis and algorithm

Given rank R and number of partition classes K, the aim of a KC-HICLAS analysis of a binary data matrix $\mathbf{D}$ is to approximate $\mathbf{D}$ as closely as possible by a binary model matrix $\mathbf{M}$, in terms of the loss function

$$L = \sum_{i=1}^{I}\sum_{j=1}^{J} |d_{ij} - m_{ij}| = \sum_{i=1}^{I}\sum_{j=1}^{J} (d_{ij} - m_{ij})^2, \tag{7}$$

and such that $\mathbf{M}$ can be represented by a (R, K) KC-HICLAS model.

The algorithm we propose successively executes two main routines. In the first routine, it looks by means of a modified version of the alternating branch-and-bound procedure for standard HICLAS analysis (Leenen and Van Mechelen, 2001) for arrays $\{\mathbf{P}, \mathbf{A}^*, \mathbf{B}^*\}$, that are such that (7) is

minimal. This routine starts from an initial configuration, $\{\mathbf{P}^{(0)}, \mathbf{B}^{*(0)}\}$ for $\mathbf{P}$ and $\mathbf{B}^*$ respectively. The initial partition matrix $\mathbf{P}^{(0)}$ is obtained by applying a K-means clustering to $\mathbf{D}$, starting from a random initialization. The initial configuration for $\mathbf{B}^*$ is obtained rationally by a built-in heuristic in the algorithm (for more details, see Leenen and Van Mechelen (2001)).

In the first routine, alternatingly three steps are further taken: (a) Conditionally upon $\mathbf{P}^{(h)}$ and $\mathbf{B}^{*(h)}$, the optimal matrix $\mathbf{A}^{*(h+1)}$ that minimizes (7) is obtained. In particular, the k^{th} row $\mathbf{A}^{*(h+1)}_{k:}$ of $\mathbf{A}^{*(h+1)}$ $(\forall k = 1, \ldots, K)$ is optimized by means of a generalized form of Boolean regression that minimizes

$$L_k = \sum_{i:p_{ik}=1} \sum_{j=1}^{J} |d_{ij} - m_{ij}| \tag{8}$$

More precisely, in this regression the values of each predictor variable are $N_k = \sum_i^I p_{ik}$ concatenated copies of each column of $\mathbf{B}^{*(h)}$, whereas the values of the criterion vector are the data entries d_{ij} $(\forall i : p_{ik} = 1; \forall j = 1, \ldots, J)$. In the next step (b), conditionally upon $\mathbf{A}^{*(h+1)}$ and $\mathbf{B}^{*(h)}$, the optimal partition matrix $\mathbf{P}^{*(h+1)}$ that minimizes (7) is obtained. This is done by estimating successively each row of $\mathbf{P}^{*(h+1)}$ by means of an exhaustive searching procedure that looks for the best assignment vector $\mathbf{x} = (0, \ldots, 1, \ldots, 0)$ of length K that minimizes

$$L_i = \sum_{j=1}^{J} |d_{ij} - \sum_{k=1}^{K} x_k m^*_{kj}| \tag{9}$$

In the last step (c), $\mathbf{B}^{*(h+1)}$ is re-estimated conditionally upon $\mathbf{A}^{*(h+1)}$ and $\mathbf{P}^{*(h+1)}$. In particular, the j^{th} row $\mathbf{B}^{*(h+1)}_{j:}$ of $\mathbf{B}^{*(h+1)}$ $(\forall j = 1, \ldots, J)$ is optimized by means of a standard Boolean regression that minimizes

$$L_j = \sum_{i=1}^{I} |d_{ij} - m_{ij}| \tag{10}$$

This procedure is repeated with $(h = 0, 1, 2, \ldots)$ until no further improvement in the loss function (7) is observed.

In the second main routine, the matrices $\mathbf{A}^*$ and $\mathbf{B}^*$ as obtained at the end of the first routine are modified such as to make them consistent with the equivalence and hierachical relations in

the model matrix $\mathbf{M}^*$ that $\mathbf{A}^*$ and $\mathbf{B}^*$ yield by (4). For this, a closure operation (Barbut and Monjardet, 1970) is successively applied to each of the two matrices $\mathbf{A}^*$ and $\mathbf{B}^*$. This operation implies that zero–entries in the two matrices are turned into one if this change does not alter $\mathbf{M}^*$ (and, hence, neither the value of the loss function (7)).

5 Simulation study

In this section, we present a simulation study in which the KC-HICLAS algorithm is evaluated with respect to sensitivity to local minima, goodness of fit (Subsection 5.2) and goodness of recovery (Subsection 5.3). In the following subsection (Subsection 5.1) the design of the simulation study is outlined.

5.1 Design and procedure

Three different types of binary $I \times J$ arrays must be distinguished in this simulation study: a true matrix $\mathbf{T}$, which is constructed by the simulation researcher and which can be perfectly represented by a KC-HICLAS model of a specific rank; a data matrix $\mathbf{D}$, which is $\mathbf{T}$ perturbed with error; and the model array $\mathbf{M}$ yielded by the KC-HICLAS algorithm.

Four aspects were systematically varied in a complete factorial design:

(a) the *Size*, $I \times J$, of $\mathbf{T}$, $\mathbf{D}$ and $\mathbf{M}$, at 6 levels: 25×15, 20×20, 80×20, 40×40, 500×50 and 150×150;

(b) the *True rank*, (R, K), of the KC-HICLAS model for $\mathbf{T}$, at 9 levels: (2,2), (2,3), (2,4), (3,3), (3,6), (3,8), (4,4), (4,10), (4,16);

(c) the *Ratio* $l : s$ of large-to-small partition class sizes, at 3 levels: .50:.50, .60:.40, .70:.30.

(d) the *Error level*, ε, which is the proportion of cells d_{ij} differing from t_{ij}, at 5 levels: .00, .05, .10, .20, .30.

All aspects will be considered random effects.

For each combination of Size $I \times J$, True rank (R, K), Ratio $l : s$ and Error level ε, 20 true matrices $\mathbf{T}$ were constructed as follows: First, a partition matrix $\mathbf{P}$ was generated by randomly assigning each of the I objects to one of the K partition classes, subject to the constraint that half of the partition classes consists of $lI/.5K$ objects and half of the classes of $sI/.5K$ objects. Subsequently, bundle matrices $\mathbf{A}^*$ and $\mathbf{B}^*$ were generated with entries that were independent realizations of a Bernoulli variable with a probability parameter chosen such that the expected proportion of ones in $\mathbf{T}$ equals 0.5, under the restiction that $\mathbf{A}^*$ and $\mathbf{B}^*$ were of Schein rank R. Finally, combining $\mathbf{P}$, $\mathbf{A}^*$ and $\mathbf{B}^*$ by the KC-HICLAS association rule yielded $\mathbf{T}$.

Next, a data matrix $\mathbf{D}$ was constructed from each true matrix $\mathbf{T}$ by altering the values of the entries chosen randomly with a probability per entry of ε. Finally, all data matrices $\mathbf{D}$ were analyzed with 100 runs of the KC-HICLAS algorithm in the True rank (R,K).

5.2 Local minima and goodness of fit

In this subsection, the KC-HICLAS algorithm is evaluated with respect to how well it succeeds in minimizing the loss function, that is, with respect to goodness of fit. To this end, the following badness-of-fit (BOF) statistic was used:

$$BOF = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (d_{ij} - m_{ij})^2}{I \times J}. \tag{11}$$

Of the 16200 analyses (5 Error levels $\times$ 3240 analyses per Error level), 6316 or 39% ended up in a solution with a BOF value larger than ε. As the Error level ε constitutes an upper bound for the BOF of the global minimum, the latter implies that the algorithm ended in a local minimum in a considerable number of cases. To investigate further the issue of local minima, we examined how many out of the 100 runs for each analysis ended in the retained solution: On average, this was the case for 37 of the 100 runs. The latter results imply that it is not unusual for the algorithm

to end in a local minimum; however, all subsequent simulation results will show that the obtained solutions are reasonably close to the underlying truth.

An analysis of variance with $BOF - \varepsilon$ as the dependent variable yielded intraclass correlations $\hat{\rho}_I$ (Haggard, 1958; Kirk, 1982) of .21 and .40 for the main effects of Size and Error Level. As the mean $BOF - \varepsilon$ across the 720 observations within each Error level are .018, .006, .001, -.010 and -.034 for ε equal to .00, .05, .10, .20, and .30, respectively (see Table 7), the main effect of error level implies that the higher ε, the easier it is for the algorithm to find a model that is as close or closer to the data as the truth is. With respect to the main effect of Size: $BOF - \varepsilon$ increases with size. Finally, the Size $\times$ Error level interaction ($\hat{\rho}_I$=.15) should be taken into account: The effect of Error level decreases with size. The other effects are not discussed: In this and the following analyses of variance only effects accounting for at least 10% of the variance of the dependent variable will be considered (i.e., $\hat{\rho}_I \geq .10$).

5.3 Goodness of recovery

In this subsection, we will evaluate how well the KC-HICLAS algorithm succeeds in recovering (1) the association relation, (2) the equivalence relations and (3) the hierarchical relations.

(1) The badness of recovery of the association relation (BOR) was assessed by the proportion of discrepancies between $\mathbf{T}$ and $\mathbf{M}$:

$$BOR = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(t_{ij} - m_{ij})^2}{I \times J}.$$

The mean BOR across the 16200 observations equals .046, which means that the model yielded by the algorithm differs on average 4.6% from the underlying truth. An analysis of variance with BOR as the dependent variable yields a main effect of Error level ($\hat{\rho}_I = .47$): except for errorfree data, badness of recovery clearly increases with higher Error levels (see Table 8). Furthermore, the Size $\times$ Error level interaction ($\hat{\rho}_I = .37$) has to be taken into account, indicating that recovery of

the truth is especially worse at combinations of high error levels and small data sizes. The latter is not too much of a surprise given that a small quantity of noisy information can be expected indeed to provide a small basis for recovering the truth.

Goodness of recovery of the association relation was also measured by the relative-recovery-gain (RRG) statistic, defined as:

$$RRG = \frac{\varepsilon - BOR}{\varepsilon},$$

RRG being undefined for $\varepsilon = 0$. This statistic equals 1 in case of perfect recovery and 0 if the model is as far from the truth as the data are. More in general, it expresses the extent to which the reconstructed data are closer to the truth than the observed data are. The mean RRG across the 12960 observations for which $\varepsilon > 0$ amounts to .738. The latter means that $\mathbf{M}$ contains, on average, 73.8% less erroneous entries than $\mathbf{D}$.

(2) With respect to the recovery of the equivalence relations, the corrected Rand index (CRI, ref) is used to compare the partition of the set of objects (resp. attributes) in the KC-HICLAS model for $\mathbf{T}$ with the partition of the set of objects (resp. attributes) in $\mathbf{M}$. This index equals 1 if the two partitions are identical and 0 if the two partitions do not correspond more than expected by chance. The mean CRI across the 16200 observations equals .764 and .811 for the objects and the attributes respectively, implying a reasonable correspondence between the true and reconstructed equivalence relations. An analysis of variance with CRI as the dependent variable yields a main effect of Error level ($\hat{\rho}_I$ amounts to .42 and .37 respectively): except for errorfree data, it holds that the higher the Error level, the lower the CRI. The main effect of Error level is qualified by a Size $\times$ Error level interaction ($\hat{\rho}_I$=.35 and .39, respectively), the latter resulting once again from the Error level effect being more important for smaller Sizes (see Table 9).

(3) To assess the recovery of the hierarchical relations, we first define the object hierarchy matrix associated with $\mathbf{M}$ as the $I \times I$ binary matrix $\mathbf{U}^{(\mathbf{M})}$, with $u_{ii'}^{(\mathbf{M})} = 1$ iff object i is hierarchically below object i' in $\mathbf{M}$. Similarly, we define an object hierarchy matrix $\mathbf{U}^{(\mathbf{T})}$ associated with $\mathbf{T}$.

Subsequently, the proportion of discrepancies between $\mathbf{U}^{(\mathrm{T})}$ and $\mathbf{U}^{(\mathrm{M})}$ was calculated yielding a badness-of-hierarchy-recovery (*BOHR*) statistic for the objects:

$$BOHR = \frac{\sum_{i=1}^{I}\sum_{i'=1}^{I}\left(u_{ii'}^{(\mathrm{T})} - u_{ii'}^{(\mathrm{M})}\right)^2}{I^2}.$$

Similarly, a *BOHR* statistic for the attributes was defined.

The mean value of *BOHR* across the 16200 observations equals .099 and .058 for the objects and the attributes respectively, implying that, on average, 90.1 and 94.2% of the true object and attribute hierarchies are recovered. Analyses of variance with *BOHR* as the dependent variable show a main effect of Error level ($\hat{\rho}_I$=.49 and .37, respectively): apart from errorfree data, recovery deteriorates with increasing Error level (see Table 10). Additionally, the Size $\times$ Error level interaction ($\hat{\rho}_I$ =.33 and .40) has to be considered: the larger the data set, the smaller the effect of Error level.

A simulation study was conducted to evaluate the KC-HICLAS algorithm. The new algorithm turned out to succeed quite well in minimizing the least-square loss function (7), as it seldom yields a model that fits the data worse than the underlying truth fits the data. Regarding the recovery of the underlying truth, we found that the models yielded by the KC-HICLAS algorithm are considerably closer to the underlying truth than the data are. This result holds for each of the three structural relations represented by the model. Moreover we found that with increasing size of the data matrix, goodness-of-fit decreased, whereas goodness-of-recovery increased. Finally, the algorithm succeeds in finding the true, when large data matrices are considered, even if the data are highly error-perturbed.

6 Two empirical applications

A KC-HICLAS analysis may be considered if one may wish (1) to limit the size of the partition of one mode in order to obtain a simpler hierarchical structure for that mode (descriptive approach),

or (2) to test the goodness of fit of a constrained a priori model of the partitioning of the mode with respect to the data (confirmatory approach). In this section, we present two illustrative applications of KC-HICLAS analysis, one of each type.

6.1 A descriptive KC-HICLAS analysis

In this first application we consider data from a study on an implicit taxonomy in psychiatric diagnosis (Van Mechelen & De Boeck, 1989). In this study, an experienced senior psychiatrist working on an intake ward of a university psychiatric clinic was asked to judge the next 30 patients arriving at his ward. First, each patient was scored by the psychiatrist on a checklist with 22 symptoms, based on the scale headings of the Psychiatric Status Schedule (Spitzer, Endicott, Fleiss, & Cohen, 1970). A symptom was scored 0 if it was absent, and 1 if it was present (in part or in whole) during the last week before the ratings. Next each patient was also given a DSM-III diagnosis by the same psychiatrist; subsequently the resulting diagnoses were coded into three (not necessarily disjoint) diagnostic categories: *affective disorder* (AD), *psychotic disorder* (PD) and *substance use disorder* (SUD), respectively.

The data matrix $\mathbf{D}$ was analyzed by means of a KC-HICLAS algorithm in ranks 1 to 6 and with K-values 1 to 9. Moreover, in order to avoid either meaningless or overparametrizated solutions, only the (R, K) pairs satisfying the condition $R \leq K \leq 2^R$ were considered (see Section 7.1 for further details). The number of discrepancies for the resulting solutions are displayed in Figure 2. To select a final solution, we made use of a two-step scree procedure: First, for each value of K we selected the optimal value of R (denoted by a dashed circle in Figure 2). Second, the optimal value of K was selected among the solutions selected in Step 1. From this procedure, we derive that the solution $(R = 3, K = 3)$ is the one to be preferred, with 13.6% discrepancies with respect to the input data $\mathbf{D}$.

Figure 3 shows the graphic representation of this solution. Three main implicit categorization

principles of the psychiatrist emerge from an inspection of the symptom structure (lower half of Figure 2). They correspond to affective problems (left lower half of Figure 2: class S1); psychotic problems (middle-right lower half of Figure 2: class S2); and substance abuse problems (rightmost symptom class: S3).

In order to examine the relationship between the implicit categorization principles and the DSM-III classifications as provided by the psychiatrist, we can calculate the proportion of patients classified in each diagnostic category with respect to the three patient classes. In class AD, 81% of the patients got a diagnosis of affective disorder. In class PD, 83% of the patients got a diagnosis of psychotic disorder. Finally in class SUD, 87% of the patients got a diagnosis of substance use disorder. The latter results appear to be consistent with the overall structure of Figure 3.

6.2 A confirmatory KC-HICLAS analysis

In this second application we present a confirmatory KC-HICLAS analysis of data from a study on archetypal psychiatric patients (Mezzich and Solomon, 1980). In this study, each of 11 psychiatrists was invited to think of a typical patient for each one of four diagnostic categories: manic-depressive/depressed (MDD), manic-depressive/manic (MDM), simple schizophrenic (SS) and paranoid schizophrenic (PS). These four diagnostic categories are part of the nomenclature of mental disorders (DSM-II) issued in 1968 by the American Psychiatric Association. The 11 psychiatrists characterized each archetypal patient by $0-6$ severity ratings on 17 symptoms from the Brief Psychiatric Rating Scale (BPRS).

For the KC-HICLAS analysis, each symptom of the original data base was trichotomized into two dummy variables indicating at least a minimal severity rating $(1-6)$ and a high severity rating $(3-6)$, respectively. This resulted in a 44×34 patient by symptom data matrix $\mathbf{D}$. Next, $\mathbf{D}$ was analyzed by means of the KC-HICLAS algorithm in ranks 2 to 4 with $K = 4$. The proportion of discrepancies for the resulting solutions were .20, .16 and .14, respectively. On the

basis of a scree test, the rank 4 solution was retained. The four archetypal patient classes of this model corresponded to the four diagnostic categories under study, with 10% misclassifications only. Figure 4 shows the graphical representation of the KC-HICLAS rank-4 solution. In order to simplify the reading of this representation we will interpret the symptom structure by mainly taking into account symptoms with a high rating.

From an a priori point of view one might expect two groups of archetypal psychiatric patients with the associated specific psychiatric symptoms: a) the schizophrenic group (S) that can be further decomposed into the SS group and the PS group, and b) the affective disorder group (A) that can further be split into the MDD group and the MDM group. However, within each group (S or A) the two subgroups share a few common symptoms only. Three other different groups emerge from the KC-HICLAS analysis. The first group (PS + MDM) is characterized by positive hypomanic affective symptoms. With this group, the paranoid schizophrenic type is further hierarchically superordinate with respect to the manic-depressive/manic type, in that PS in addition is also characterized by positive psychotic signs (such as hallucinations). The second group (MDD) is characterized by depressive affective symptoms (S2, S7). Finally the third group (SS) is represented by the negative psychotic sign blunted affect.

7 Discussion and possible extensions

In this section, we will first discuss the link between the KC-HICLAS model rank R and the upper bound K. Next, we will explore the relationships between KC-HICLAS and standard partitioning techniques that minimize certain clustering criteria, such as standard K-means clustering as well as two–way two–mode partitioning techniques. Finally, two possible extensions of KC-HICLAS will be proposed.

7.1 Relationship between the model ranks R and K

In general the model rank R specifies the *global complexity* of a HICLAS model. This global complexity implies an upper–bound for the number of object (resp. attribute) classes, that is, 2^R. From this it follows that in a KC-HICLAS analysis it is meaningless to select a value of the maximal number of object classes K greater than the general upper bound 2^R, given that this would imply that at least $K - 2^R$ object classes would show up to be equivalent. On the other hand, if in KC-HICLAS one would specify a value of $K < R$, then an overparametrization of the model would occur. Indeed, assume that $\mathbf{M} = \mathbf{P}\mathbf{M}^*$, where $\mathbf{M}^* = \mathbf{A}_R^* \otimes \mathbf{B}_R^*$ denotes a rank-R model decomposition of $\mathbf{M}^*$. Now the smallest integer R' which is such that $\mathbf{M}^*$ can be given a rank-R' decomposition (i.e., the Schein rank of $\mathbf{M}^*$), is necessarily less than or equal to the number of rows of $\mathbf{M}^*$ (Kim, 1982), that is K; hence, $R' \leq K$. From $K < R$, then it further follows that $R' < R$, which implies that a rank-R decomposition of $\mathbf{M}$ boils down to an overparametrization. All of the above can be summarized as

$$R \leq K \leq 2^R \tag{12}$$

7.2 Relationship with other models

As a partitioning model for two-way two-mode data, KC-HICLAS is similar to standard partitioning techniques that minimize a certain clustering criterion, such as, for example, standard one-mode K-means (MacQueen, 1967) as well as several two-mode partitioning techniques (e.g. Bock, 2003; Castillo and Trejos, 2002; Govaert, 1995). In general, however the specific optimization criteria used in K-means clustering and KC-HICLAS are slightly different. In particular, standard K-means produces an optimal K–partition of the objects by minimizing the trace of the within group sum–of–squares and crossproducts matrix $\mathbf{W}$, or, equivalently, by minimizing the sum of the squared Euclidean distances between objects and their cluster mean; from its part KC-HICLAS seeks for an optimal K–partition of the objects by minimizing the least absolute deviation or least squares

loss function as defined in (7), which further comes down to minimizing the summed within-group squared Euclidean distances between the objects and their respective partition class centroids. From all this it can be derived that the optimization in K-means clustering and KC-HICLAS differ with regard to how the within group heterogeneity is quantified. In K-means clustering this is done in terms of squared Euclidean distances to the cluster means, in KC-HICLAS this is done in terms of squared Euclidean distances to the partition class centroids, the latter as defined in Section 3, being a binary vectors that can be considered the Boolean counterpart of the real-valued means.

Furthermore, KC-HICLAS is also related to factorial K-means analysis for two-way real data (Vichi and Kiers, 2001): Factorial K-means is characterized by the combination of a discrete clustering model and a continuous factorial model that are fitted simultaneously to two-way real data. The objective of a factorial K-means analysis is to identify the optimal partition of one of the two modes, optimality being defined in terms of a least squares criterion with regard to centroids in a space of reduced dimension. Like factorial K-means also KC-HICLAS restricts the number of partition classes ($\leq K$) of one of the two modes with, in particular, the partition class centroids being constrained to be located in a Boolean vector space that is reduced in terms of the model rank R. However, like the standard K-means method, factorial K-means does not allow for any additional structural organization of the partition classes implied by the model.

7.3 Possible extensions

Two possible extensions of the K-Centroids hierarchical classes model may be considered: First, in the present paper, KC-HICLAS has been proposed as a model for two-way two-mode Boolean data. However, the current approach can be straightforwardly extended to rating-valued data. In particular, a KC-HICLAS model for rating data could be considered as a particular constrained instance of the HICLAS-R model family (Van Mechelen, Lombardi and Ceulemans, 2002). In a rating-valued context, the approximation problem would boil down to seeking a KC-HICLAS

rating-valued model matrix $\mathbf{M}$ that approximates the rating-valued data matrix $\mathbf{D}$ such that (7) is minimal. A second further extension of the KC-HICLAS model could imply an explicit limit on the classification of the elements of both modes to at most K_1 (resp. K_2) partition classes. In this extension, the association rule (4) would change into

$$\mathbf{M} = \mathbf{P}(\mathbf{A}^* \otimes \mathbf{B}^{*\prime})\mathbf{Q} \tag{13}$$

with $\mathbf{Q}$ being the partition matrix of the second mode.

References

BAIER, D., GAUL, W., and 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). Heidelberg: Springer-Verlag.

BARBUT, M., and MONJARDET, B. (1970), *Ordre et classification: Algèbre et combinatoire* (2 Vols.) Hachette, Paris.

BOCK, H-H. (2003), Two-way clustering for contingency tables: maximizing a dependence measure. In M. Schader, W. Gaul, & M. Vichi (Eds.), *Between data science and applied data analysis* (pp. 143–155). Heidelberg: Springer-Verlag.

CASTILLO, W., and TREJOS, J. (2002), Two-mode partitioning: review of methods and application of tabu search. In K. Jajuga, A. Sokołowski & H.-H. Bock (Eds.), *Classification, clustering, and data analysis* (pp. 43–51). Heidelberg: Springer-Verlag.

CEULEMANS, E., and VAN MECHELEN, I. (in press). Tucker2 hierarchical classes analysis. *Psychometrika.*

CEULEMANS, E., VAN MECHELEN, I., and LEENEN, I. (in press). Tucker3 hierarchical classes analysis. *Psychometrika.*

DE BOECK, P., and ROSENBERG, S. (1988), Hierarchical classes: Model and data analysis. *Psychometrika, 53, 361–381.*

ECKES, T. (1991), Bimodale Clusteranalye: Methoden zur Klassifikation von Elementen zweier Mengen. *Zeitschrift für experimentelle und angewande Psychologie, XXXVIII, 201–225.*

ECKES, T., and ORLIK, P. (1993), An error variance approach to two-mode hierarchical clustering. *Journal of Classification, 10, 51–74.*

GAUL, W., and SCHADER, M. (1996), A new algorithm for two-mode clustering. In H.-H. Bock & W. Polasek (Eds.), *Data analysis and information systems* (pp. 15–23). Heidelberg: Springer-Verlag.

GETZ, G., LEVINE, E., and DOMANY, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Prooceding of the National Academy of Science, Vol. 97, no. 22* 12079-12084.

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

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

KIM, K.H. (1982), *Boolean matrix theory.* Marcel Dekker, New York.

KIRK, R.E. (1982), *Experimental design: procedures for the behavioral sciences* (2nd ed.). Brooks/Cole, Belmont, CA.

LEENEN, I., and VAN MECHELEN, I. (2001), An evaluation of two algorithms for hierarchical classes analysis. *Journal of Classification, 18, 57–80.*

LEENEN, I., VAN MECHELEN, I., DE BOECK, P., and ROSENBERG, S. (1999), INDCLAS: A three-way hierrachical classes model. *Psychometrika, 64*, 9-24.

MACQUEEN, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281-297.

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

SPITZER, R.L., ENDICOTT, J., FLEISS, J.L., and COHEN, J. (1970). The Psychiatric Status Schedule: A technique for evaluating psychopatology and impairement in role functioning. *Archives of General Psychiatry, 23*, 41-55.

VAN MECHELEN, I., BOCK, H-H., and DE BOECK, P. (2003), Two-mode clustering: a structured overview. Submitted for publication.

VAN MECHELEN, I., and DE BOECK, P. (1989). Implicit taxonomy in psychiatric diagnosis: a case study. *Journal of Social and Clinical Psychology, 8*, 276-287.

VAN MECHELEN, I., DE BOECK, P., and ROSENBERG, S. (1995), The conjunctive model of hierarchical classes. *Psychometrika, 60, 505–521.*

VAN MECHELEN, I., LOMBARDI, L., and CEULEMANS, E. (2002), Hierarchical classes modeling of rating data. Submitted for publication.

VICHI, M., and KIERS, H.A.L. (2001), Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis, 37, 49–64.*

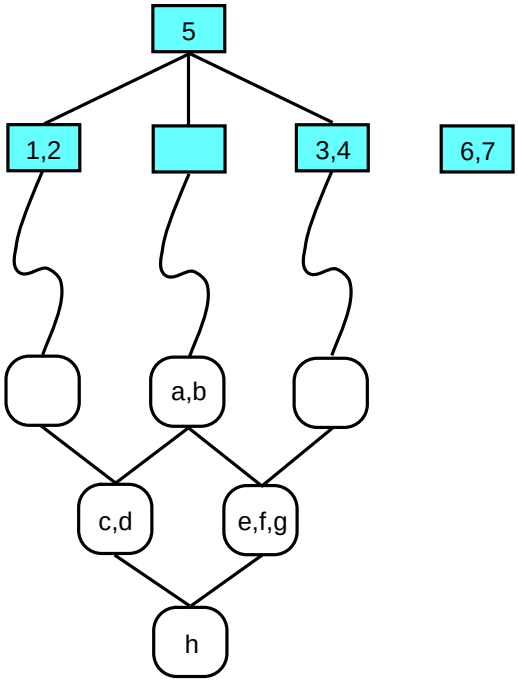


Figure 1: Graphic representation of the hierarchical classes model of Table 3. Empty boxes denote empty classes.

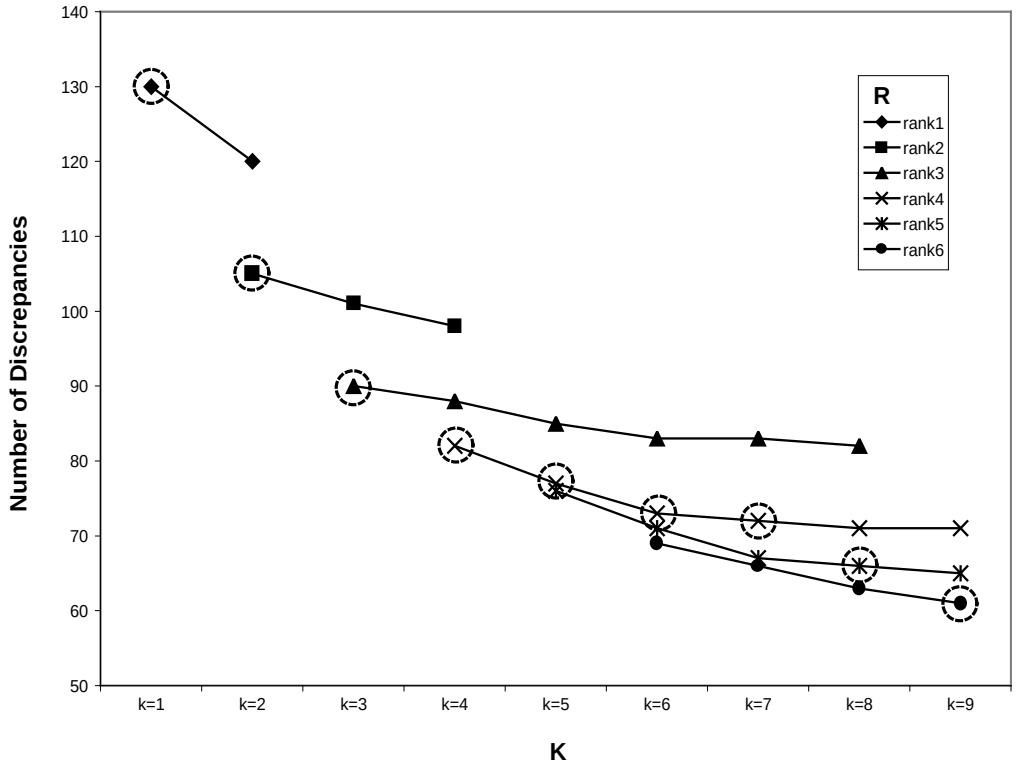


Figure 2: Scree plot of number of discrepancies as a function of the model rank R and the number of clusters K. Circled points denote the preferred solution within each of the K-values.

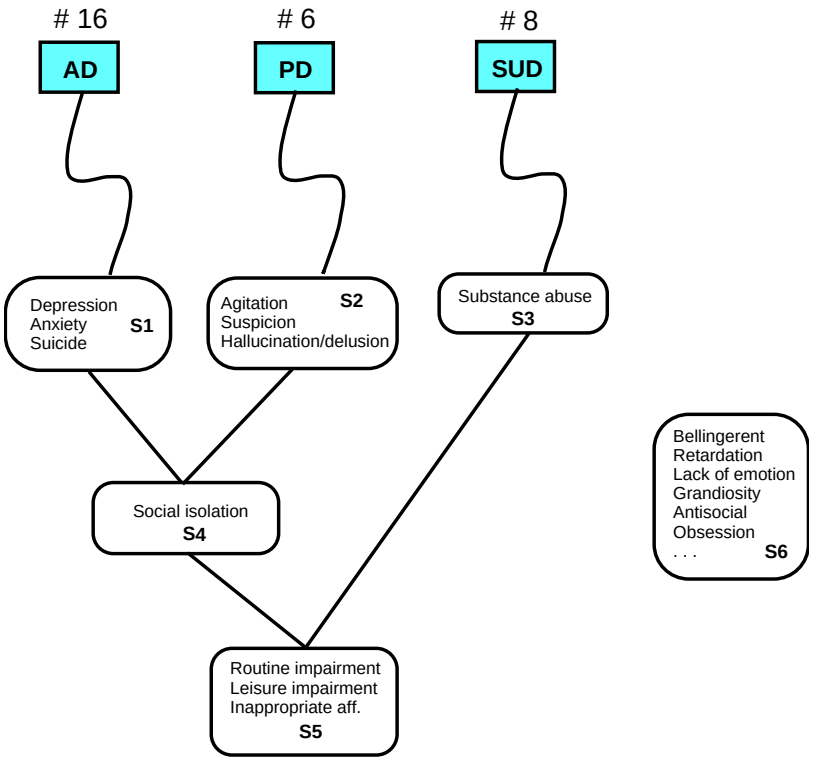


Figure 3: Graphic representation of the KC-HICLAS rank-3 solution with $K = 3$. Patient classes (resp. symptoms) are displayed in the upper half (resp. lower half) of the figure.

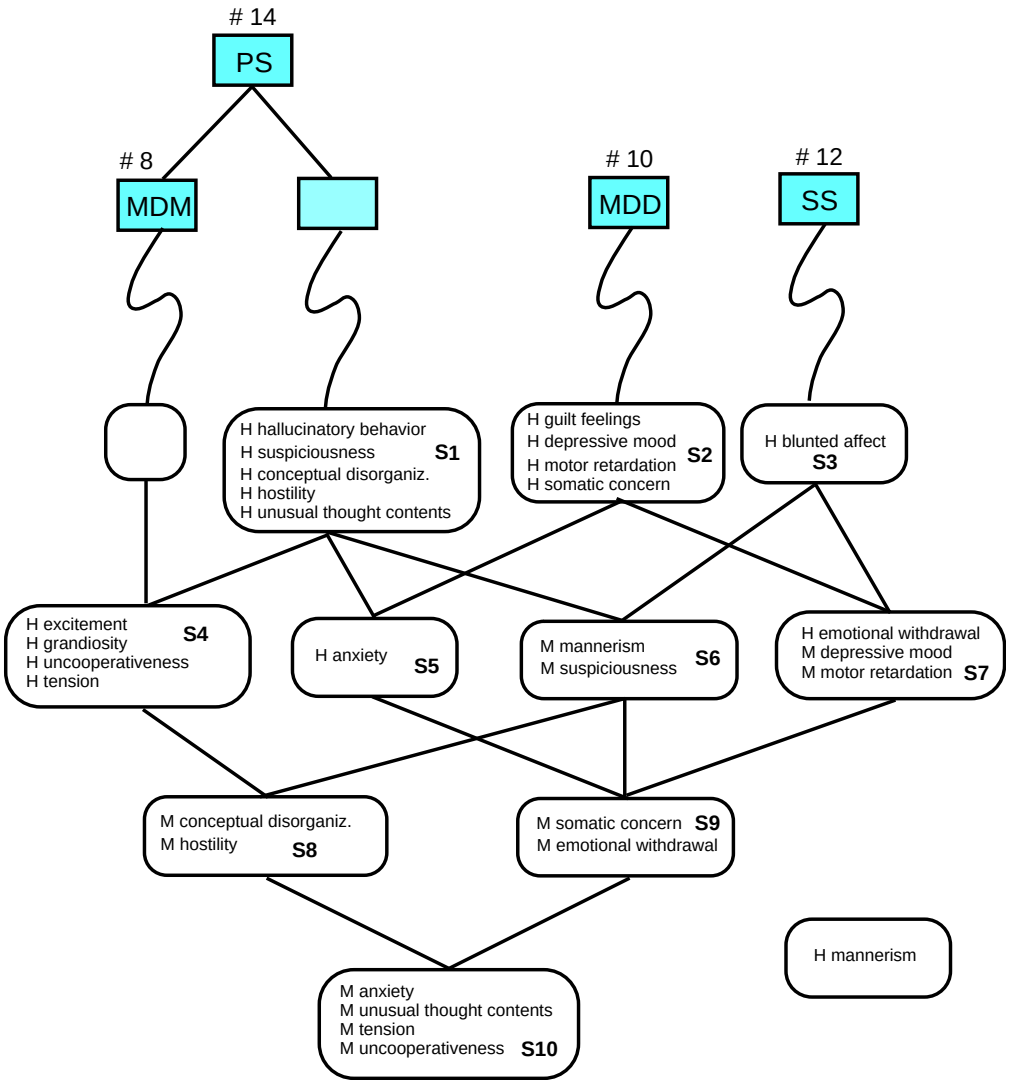


Figure 4: Graphic representation of the KC-HICLAS rank ($R = 4, K = 4$) solution with $K = 4$. Diagnostic categories (resp. symptoms) are displayed in the upper half (resp. lower half) of the graphic. M and H denote medium rating and high rating, respectively.

Table 1: Hypothetical Two-Way Two-Mode Binary Data Matrix **D**

	a	b	c	d	e	f	g	h
Obj 1	0	0	1	1	0	0	0	0
Obj 2	0	0	1	1	0	0	0	1
Obj 3	1	0	0	0	1	1	1	1
Obj 4	0	1	0	0	1	1	1	1
Obj 5	1	1	1	1	0	1	1	1
Obj 6	0	0	0	0	0	0	0	0
Obj 7	0	0	0	0	0	0	0	0

Table 2: Model matrix **M** of the data matrix **D** in Table 1

	a	b	c	d	e	f	g	h
Obj 1	0	0	1	1	0	0	0	1
Obj 2	0	0	1	1	0	0	0	1
Obj 3	0	0	0	0	1	1	1	1
Obj 4	0	0	0	0	1	1	1	1
Obj 5	1	1	1	1	1	1	1	1
Obj 6	0	0	0	0	0	0	0	0
Obj 7	0	0	0	0	0	0	0	0

Table 3: Disjunctive Hierarchical Classes Model for the Matrix in Table 2

	Row Bundles				Column Bundles		
Row Entries	I	II	III	Column Entries	I	II	III
Obj 1	1	0	0	a	0	1	0
Obj 2	1	0	0	b	0	1	0
Obj 3	0	0	1	c	1	1	0
Obj 4	0	0	1	d	1	1	0
Obj 5	1	1	1	e	0	1	1
Obj 6	0	0	0	f	0	1	1
Obj 7	0	0	0	g	0	1	1
				h	1	1	1

Table 4: Hypothetical data matrix **D** and related KC-HICLAS model matrix **M**

D

	Attributes			
Objects	a	b	c	d
Obj 1	1	0	1	0
Obj 2	1	1	1	0
Obj 3	1	0	1	0
Obj 4	0	0	1	1
Obj 5	0	0	1	0
Obj 6	0	1	1	0
Obj 7	0	0	1	0

M

	Attributes			
Objects	a	b	c	d
Obj 1	1	0	1	0
Obj 2	1	0	1	0
Obj 3	1	0	1	0
Obj 4	0	0	1	0
Obj 5	0	0	1	0
Obj 6	0	1	1	0
Obj 7	0	1	1	0

Table 5: Centroids decomposition for matrix **M** in Table 4

P

	Centroids		
Objects	α	β	γ
Obj 1	1	0	0
Obj 2	1	0	0
Obj 3	1	0	0
Obj 4	0	1	0
Obj 5	0	1	0
Obj 6	0	0	1
Obj 7	0	0	1

M*

	Attributes			
Centroids	a	b	c	d
α	1	0	1	0
β	0	0	1	0
γ	0	1	1	0

Table 6: Model decomposition for matrix $\mathbf{M}^*$ in Table 5

$\mathbf{A}^*$

	Bundles		
Centroids	I	II	III
α	1	1	0
β	1	0	0
γ	1	0	1

$\mathbf{B}^*$

	Bundles		
Attributes	I	II	III
a	0	1	0
b	0	0	1
c	1	1	1
d	0	0	0

Table 7: Mean Differences between Badness of Fit and ε at Levels of Size $\times$ Error

	Error Level					
Size	.00	.05	.10	.20	.30	Overall
25 $\times$ 15	.004	-.000	-.006	-.027	-.073	-.020
20 $\times$ 20	.004	.000	-.004	-.025	-.071	-.019
80 $\times$ 20	.013	.003	-.000	-.008	-.032	-.005
40 $\times$ 40	.007	.002	.001	-.004	-.025	-.004
500 $\times$ 50	.053	.022	.012	.003	-.003	.017
150 $\times$ 150	.030	.006	.003	.002	.001	.008
Overall	.018	.006	.001	-.010	-.034	-.004

Table 8: Mean Badness of Recovery at Levels of Size × Error

	Error Level					
Size	.00	.05	.10	.20	.30	Overall
25 × 15	.004	.010	.025	.101	.240	.076
20 × 20	.004	.008	.021	.100	.243	.076
80 × 20	.013	.006	.010	.046	.158	.047
40 × 40	.007	.003	.005	.030	.154	.040
500 × 50	.053	.025	.016	.011	.041	.029
150 × 150	.030	.007	.004	.004	.015	.012
Overall	.018	.010	.014	.049	.142	.046

Table 9: Mean Corrected Rand Index at Levels of Size × Error

	Objects						Attributes					
Size	.00	.05	.10	.20	.30	Overall	.00	.05	.10	.20	.30	Overall
25 × 15	.946	.883	.781	.469	.166	.649	.971	.927	.838	.552	.235	.705
20 × 20	.963	.919	.844	.526	.185	.688	.970	.920	.831	.522	.207	.690
80 × 20	.889	.908	.876	.654	.283	.722	.924	956	.953	.858	.504	.839
40 × 40	.952	.974	.961	.832	.419	.828	.960	.968	.943	.790	.410	.814
500 × 50	.646	.807	.861	.885	.679	.776	.759	.884	.921	.962	.932	.892
150 × 150	.795	.931	.969	.978	.927	.920	.858	.959	.967	.965	.893	.928
Overall	.865	.904	.882	.724	.443	.764	.907	.936	.909	.775	.530	.811

Table 10: Mean Badness of Hierarchy Recovery at Levels of Size × Error

	Objects						Attributes					
Size	.00	.05	.10	.20	.30	Overall	.00	.05	.10	.20	.30	Overall
25 × 15	.035	.049	.084	.205	.341	.143	.009	.023	.046	.130	.248	.091
20 × 20	.024	.038	.066	.186	.328	.128	.009	.022	.046	.142	.266	.097
80 × 20	.045	.038	.050	.132	.295	.112	.021	.014	.015	.040	.148	.048
40 × 40	.023	.016	.024	.075	.240	.076	.013	.010	.017	.059	.183	.056
500 × 50	.134	.067	.052	.052	.131	.087	.077	.036	.023	.012	.022	.034
150 × 150	.076	.035	.038	.033	.047	.046	.045	.014	.011	.012	.031	.023
Overall	.056	.041	.052	.114	.230	.099	.029	.020	.026	.066	.150	.058