Statistical Methods in Medical Research 2004; **13**: 363–394

Two-mode clustering methods: a structured overview

Iven Van Mechelen, KU Leuven, Belgium, **Hans-Hermann Bock**, RWTH, Aachen, Germany and **Paul De Boeck**, KU Leuven, Belgium

In this paper we present a structured overview of methods for two-mode clustering, that is, methods that provide a simultaneous clustering of the rows and columns of a rectangular data matrix. Key structuring principles include the nature of row, column and data clusters and the type of model structure or associated loss function. We illustrate with analyses of symptom data on archetypal psychiatric patients.

1 Introduction

Statistical methods, explorative or inferential, have to deal with observed or recorded data. These data often take the form of a rectangular table, that is, an I by J data matrix $\mathbf{X} = (x_{ij})$. The entries x_{ij} of $\mathbf{X}$ may refer to the values of J variables $j = 1, \ldots, J$ measured for each of I cases (objects, persons) $i = 1, \ldots,$ I; alternatively they may refer to the values of a criterion variable recorded as a function of two categorical predictor variables with I and J categories, respectively, or to proximities between the elements of two sets with I and J elements, respectively. Given such a data table, one may wish to retrieve its underlying structure in various respects, for example, for detecting dependencies, trends, and so on. This paper is devoted to methods that aim to reveal such structures in terms of suitable joint clusterings of the rows and the columns of the data matrix.

Literature provides a plethora of clustering methods for data under various data assumptions, with different clustering structures and homogeneity concepts, without or with some optimization criterion (e.g., Bock,[1] Hartigan,[2] Arabie *et al.*,[3] Everitt *et al.*,[4] Jain and Dubes,[5] Celeux *et al.*,[6] Kaufman and Rousseeuw,[7] Mirkin,[8] Gordon[9] and the other papers in this issue). Therefore we could use any of these methods for clustering the rows (objects, say) of the data table; moreover, the same methods could be used to obtain a clustering of the columns (variables, say). If we are now interested in a clustering of both the rows and the columns (objects and variables) we can perform cluster analyses of both types of entities successively (and independently) (for an early example see Tyron[10]). As an alternative to the latter approach, one may wish to obtain clusterings of rows and columns *simultaneously* rather than successively.[11] It is the latter type of strategy that constitutes the focus of the present paper.

The references listed subsequently show that two-mode clustering has been widely applied in a broad range of domains including marketing, psychology and sociology. Within the domains of microbiology and medicine, quite recently important

Address for correspondence: Iven Van Mechelen, Psychology Department, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium. E-mail: iven.vanmechelen@psy.kuleuven.ac.be

10.1191/0962280204sm373ra

applications have appeared with respect to the study of microarrays for DNA analysis. In particular with regard to the latter, with the advent of new tools from microbiology, it is possible to measure the amount of a 'gene' (a small part of a DNA strain) that is present in some tissue or in a biological material under investigation. Such measurements are typically obtained for a large number I of genes i (typically I is about 10 000) and a moderate number J of tissues j (situations, patients, time points and so on) by using microchips with I 'pixels' or 'spots', one for each gene. Corresponding 'expression data' are compiled in the data matrix $\mathbf{X} = (x_{ij})_{I \times J}$. Then a major problem (e.g., in cancer or survival research) consists of clustering the genes (rows) with a view to obtain 'homogeneous' (e.g., functionally equivalent) gene classes, and in clustering the patients or tissues (columns) in order to obtain groups of similarly reacting patients or tissues, typically also with the idea to predicting the patient's behavior from a class of his gene configuration. In this context, simultaneous clustering methods have been broadly applied in the recent past, with adaptations to the special situation of DNA analysis (with, e.g., dependence among the genes or rows of $\mathbf{X}$). For example, refer to Getz *et al.*,[12] Li and Zha,[13] Pollard and van der Laan[14,15] and Jörnsten and Yu.[16]

As advantages of a simultaneous rather than a sequential approach to the clustering of rows and columns, two key elements can be considered. First, the mathematical structures or models as implied by several simultaneous clustering methods cannot be reduced to a simple concatenation or addition of constituent row and column clusterings (e.g., by allowing for different column clusterings in distinct row clusters). Related to the former, several simultaneous clustering methods imply the optimization of an overall objective function that cannot be reduced to a simple combination of constituent row and column objective functions. Secondly, and more importantly, a simultaneous clustering may highlight the association between the row and column clusterings that appear as linked clusterings from the data analysis (which may, e.g., considerably facilitate their characterization). Alternatively, one might say that simultaneous approaches allow the researcher to characterize the nature of the interaction or of the dependence structure between row and columns, as implied by the data.

Starting with the pioneering work of Hartigan,[2,17] who significantly contributed to the development of the simultaneous clustering domain both conceptually and algorithmically, and with some decision-theoretical investigations by Bock,[18] a broad range of simultaneous clustering methods has been developed by various authors. In the past, a few attempts have been made to structure the whole of the resulting methods,[19–22] but a comprehensive overview of the domain is still lacking, and taxonomic efforts in this area have been criticized.[23] Methods for simultaneous clustering seem very heterogeneous, both in terms of underlying mathematical structures and models, and in terms of principles and tools used in the associated data analysis. As a consequence, the simultaneous clustering domain has never been easily accessible.

The principal aim of the present paper is to provide the reader with a *taxonomic overview* of the methods in question. The emphasis in this overview will be on the characterization of the *mathematical structures* or *models* underlying the distinct methods and on the *objective* or *loss functions* that are optimized by the associated algorithms or heuristics. We will avoid a discussion of the actual algorithms or algorithmic technicalities and of more advanced issues such as goodness-of-fit measures and model selection. We will further limit our overview to extant methods (although, as will become clear from our exposition subsequently, there are still many empty holes left in the taxonomy of simultaneous clustering methods).

The structure of the remainder of this paper is as follows: Section 2 will introduce a number of conceptual preliminaries. Section 3 will present the actual taxonomy, and in Section 4 we will illustrate three types of methods with analyses of symptom data on archetypal psychiatric patients. Section 5 will present some concluding remarks.

2 Conceptual preliminaries

In this section we will introduce a number of important conceptual preliminaries that will play a key role as the major structuring principles of the taxonomy of methods and approaches we will outline. Some of these conceptual distinctions will also turn out to be highly relevant for the selection of a suitable simultaneous clustering method for a given data set and a theoretical problem at hand.

2.1 Data

Data matrices are recorded in different contexts, and their rows, columns and entries can have different conceptual structures. In order to typify the various cases, Carroll and Arabie[24] have introduced some terminology (which in turn relies on work by Tucker[25]). To use this terminology, a data set is conceived as a mapping x from a Cartesian product $\mathrm{S} = \mathrm{S}_1 \times \mathrm{S}_2 \times \ldots \times S_n$ of n sets $\mathrm{S}_1, \ldots, \mathrm{S}_n$ to some (typically univariate) domain Y: for any n-tuple $(s_1, s_2, \ldots, s_n)$ with $s_1 \in \mathrm{S}_1, \ldots, s_n \in \mathrm{S}_n$ a value $x(s_1, s_2, \ldots, s_n)$ from Y is recorded. The total number n of constituent (possibly identical) sets of S is called the number of *ways* in the data, whereas the number of *distinct* sets in S is called the number of *modes*. Therefore, any data set that can be naturally written in matrix form is called two-way ($n = 2$); if the rows and columns of a two-way data set refer to distinct sets of entities ($\mathrm{S}_1 \neq \mathrm{S}_2$), the data set is called two-mode. In the remainder of this paper we will exclusively focus on *two-way two-mode data*. Furthermore, the two sets involved in the data will be denoted by $\mathcal{R}$ (row mode) and $\mathcal{C}$ (column mode), and will be assumed to contain I and J elements, respectively. Recall also that the data entries will be denoted by $\mathbf{X} = (x_{ij})$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.

Motivated by practical cases, we will distinguish within the family of two-way two-mode data among three types:

1) *Case by variable type*: The two modes refer to cases and variables, for example, patients and symptoms; case by variable data can be binary, polytomous or continuous.
2) *Categorical predictor type*: The row and column modes refer to the categories of two (categorical) predictor variables U and V, and the data entries x_{ij} refer to the values of some single criterion variable; a special case of this data type is seen if the x_{ij}s are the observed joint frequencies of the categories i and j of the variables U and V – in this case the matrix $\mathbf{X}$ boils down to a contingency table.
3) *Proximity type*: The data entries are considered as two-mode proximities, that is, x_{ij} pertains to the similarity or dissimilarity of row element i and column element j. Note that in such a proximity interpretation, the data can be considered incomplete

if $\mathcal{R} \neq \mathcal{C}$: a full set of proximities would pertain to all $(I+J)(I+J-1)/2$ pairs in $\mathcal{R} \cup \mathcal{C}$, whereas proximities within $\mathcal{R}$ and within $\mathcal{C}$ are missing in the data. Note also that, while the idea of closeness may be intuitively clear, formal criteria to define the concept of (one-mode or two-mode) proximities are lacking. As such, case by variable data (type 1) can usually be reinterpreted as proximities (type 3); moreover, the reverse also holds.

A further important distinction pertaining to the data refers to the question as to whether data entries are comparable within each data column only, within each data row only or across the full data matrix.[17] Carroll and Arabie[24] use in this regard the terms of *column-conditional, row-conditional*, and *matrix-conditional* data (the matrix-conditional in the present case being identical to *unconditional* data). Column-conditional data, for instance, occur if one considers case by variable data with variables being measured on different scales (e.g., variable 1 in mmHg and variable 2 in kg).

2.2 Clusters, classification structures and data analysis

2.2.1 Meaning of the term cluster

In this paper, a *cluster* E within a set S will typically be any subset of S (useful clusters will have some homogeneity properties). Such clusters (*crisp* clusters) have a clear cut (0/1) membership function: A cluster E can be formalized in terms of a binary membership vector $(e_1, \ldots, e_i, \ldots, e_I)$, with $I = \#\mathrm{S}$, and with $e_i = 1$ iff element i belongs to E and $e_i = 0$ otherwise. The restriction to crisp clusters links up with the philosophical tradition of Frege,[26] and implies that we will leave aside fuzzy clustering approaches where the membership value of an element i of set S to cluster E may be any element in the interval [0,1]. [One may note that in some stochastic approaches to (one-mode and two-mode) clustering crisp clusters are assumed on the basic level of the underlying model only; in such cases we have a *latent* crisp cluster membership, whereas the data analyst's knowledge of it will typically be fuzzy – in terms of prior or posterior probabilities.]

Quite generally, we define a *clustering* or *classification* of a set S as a system $(\mathrm{E}_1, \ldots, E_h, \ldots, E_H)$ of subsets of S. If $I = \#\mathrm{S}$, such a clustering can be described by a binary $I \times H$ membership matrix $\mathbf{E} = (e_{ih})$ with $e_{ih} = 1$ iff element i belongs to the hth cluster E_h and $e_{ih} = 0$ otherwise.

2.2.2 Elements and set theoretical nature of the clustering

Two important distinctions are to be made with regard to the kind of clusters. A first distinction pertains to the *elements of the clusters* E (see Hartigan[17]): these can be *rows* (yielding row clusters $\mathrm{E} \subseteq \mathcal{R}$), *columns* (yielding column clusters $\mathrm{E} \subseteq \mathcal{C}$) and *cells in the data matrix* (yielding data clusters $\mathrm{E} \subseteq \mathcal{R} \times \mathcal{C}$); data clusters are further restricted to Cartesian products $\mathrm{E} = \mathrm{F} \times \mathrm{G}$ of a row cluster $\mathrm{F} \subseteq \mathcal{R}$ and a column cluster $\mathrm{G} \subseteq \mathcal{C}$. Note that the different clusters should not necessarily consist of contiguous rows, columns or data cells. Note further that, while a data cluster $\mathrm{E} = \mathrm{F} \times \mathrm{G}$ always implies a row cluster F and a column cluster G, in several two-mode clustering methods *the reverse does not hold* (i.e., not every pair F and G of a row and column cluster obtained by the clustering method implies a valid data cluster $\mathrm{F} \times \mathrm{G}$). In terms of notation, we will further denote the row clusters implied by a given clustering method by ($\mathrm{A}_1, \ldots, A_r, \ldots, \mathrm{A}_R$), the column clusters by ($\mathrm{B}_1, \ldots, \mathrm{B}_c, \ldots, \mathrm{B}_C$) and the data clusters by ($\mathrm{E}_1, \ldots, \mathrm{E}_d, \ldots, \mathrm{E}_D$); the row clustering will further be described by the $I \times R$ membership matrix $\mathbf{A} = (a_{ir})$ and the column clustering by the $J \times C$ membership matrix $\mathbf{B} = (b_{jc})$.

A second distinction pertains to the *set-theoretical relations* between the different clusters of a classification. In this regard we will distinguish between: 1) partitions, 2) nested clusterings and 3) overlapping clusterings. Hypothetical membership matrices for the three types of clustering are displayed in Table 1.

1) Partitions consist of a certain number of non-empty, nonintersecting clusters that span the full set under consideration (i.e., the full set of I rows, J columns, or $I \times J$ data cells, respectively).
2) Nested clusterings do include intersecting clusters; in case of a non-empty intersection, however, one of the intersecting clusters should necessarily be a subset of the other; note that a *hierarchical clustering* on a set is an important special case of a nested clustering that includes both the set itself and all singletons corresponding to the elements of the set.
3) Overlapping clusterings finally include intersecting, non-nested clusters.

Three important remarks are further to be made with regard to the two distinctions mentioned earlier:

1) In most cases we will apply the distinction between partitions, nested and overlapping clusterings only to the end result of algorithmic processes. Many cluster algorithms proceed in a stepwise fashion, indeed; as such, the end goal of some stepwise procedure may, for instance, be some partition that is obtained by a series of successive splits; if both the intermediate and final results of such a stepwise splitting procedure would be taken into account, the full set of clusters would necessarily be a nested one; if in such a case, however, the intermediate results are of no final importance, we will consider the structure under study a simple partition.
2) In case of a nested or overlapping clustering ($\mathrm{E}_1, \ldots, \mathrm{E}_h, \ldots, \mathrm{E}_H$) on a set S (with $\#\mathrm{S} = I$), one may always derive from the $I \times H$ membership matrix E of such a clustering a partition of S by grouping together all entities $i \in \mathrm{S}$ with the same

Table 1 Hypothetical membership matrices for a partitioning, a nested clustering and an overlapping clustering of the set S = {a,b,c,d,e,f}

	Clusters of a partitioning				Clusters of a nested clustering				Clusters of an overlapping clustering			
Entities	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
a	1	0	0	0	1	0	1	1	1	0	0	0
b	1	0	0	0	1	0	1	1	1	0	0	1
c	0	1	0	0	0	0	1	1	1	1	0	1
d	0	1	0	0	0	1	0	1	1	1	0	1
e	0	0	0	1	0	1	0	1	0	1	0	1
f	0	0	1	0	0	1	0	1	0	0	1	0

membership pattern $(e_{i1}, \ldots, e_{ih}, \ldots, e_{iH})$ in **E**. As an example, from the rightmost membership matrix in Table 1, one may derive the partition {{a}, {b}, {c,d}, {e}, {f}} of the underlying set {a,b,c,d,e,f}.

3) The distinctions 'type of the cluster elements' (rows, columns and data cells) and 'set-theoretical nature of the clustering' (partition, nested clustering and overlapping clustering) are to be considered simultaneously, since a data clustering and its implied row and column clusterings may be of different types. An extreme example is displayed in Figure 1, with row, column and data clusterings of three different types.

2.2.3 Symmetry

Simultaneous clustering methods can be characterized in terms of whether they deal in a symmetric or an asymmetric way with the two modes involved in the data. For asymmetric methods (which generally may be considered an underinvestigated topic), two variants can further be obtained by interchanging the roles of the two modes under study. Viewed from the perspective of case by variable data, the symmetry/asymmetry aspect touches upon a discussion within the one-mode clustering domain as to whether different kinds of methods should be considered for a one-mode clustering of cases and a one-mode clustering of variables. In this regard, authors such as Braverman[27] and Hartigan[2] have advocated the use of qualitatively different approaches for both types of clusterings. For instance, for clusterings of variables unlike for clusterings of cases, criteria implying high within-cluster correlations as opposed to low or zero between-cluster correlations have been proposed. The issue as to whether an asymmetric rather

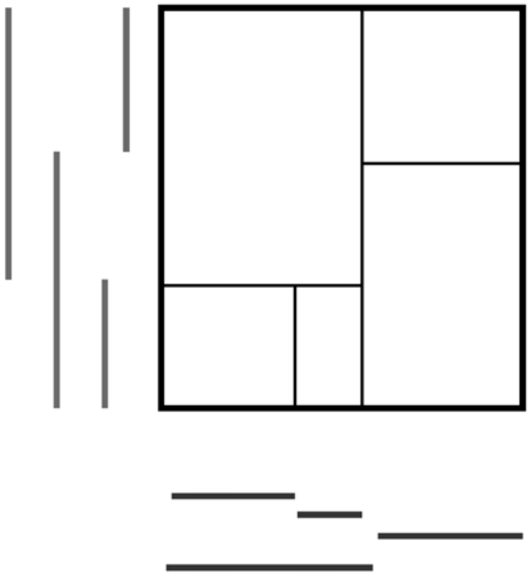

Figure 1 Hypothetical example of a data partitioning yielding an overlapping row clustering and a nested column clustering.

than a symmetric two-mode clustering method is most appropriate for the analysis of a data set further also directly relates to the conditionality of the data; clearly, symmetric approaches typically assume matrix-conditionality.

2.2.4 Level of modeling and optimization

A final important distinction between clustering methods pertains to the level of sophistication of modeling and optimization. Three levels may be discerned:

1) *Procedural level* (no optimization criterion or no model): On a first level, one may locate various computational procedures or clustering algorithms that do not imply an overall loss or objective function to be optimized, or no explicit mathematical model structure that is fitted to the data; often, such approaches use in a stepwise fashion some heuristic rule that at best indirectly relates to some overall criterion and/or some implicit mathematical structure.
2) *Deterministic level* (optimization criterion based on a deterministic model): On a second level, one may locate procedures that aim at optimizing some overall loss function (albeit sometimes only in a heuristic way) that is implied by the fitting of some deterministic mathematical structure to the data.
3) *Stochastic level* (optimization criterion based on a stochastic model): On a third level, one may locate procedures that aim at optimizing a global criterion on the basis of a stochastic model that is fitted to the data, that is, a model implying distributional assumptions. Approaches to this third level are the most complete ones in that they include a full account of the relationship between the core mathematical structure of the model and the data.

Within the stochastic level, a further distinction proves to be adequate, namely one between 3a) fixed-partition and 3b) random-partition clustering approaches[28]:

3a) Fixed-partition clustering methods assume an underlying deterministic clustering as well as parametric within-cluster distributional specifications. Optionally, they may in addition also include assumptions on the plausibility of the deterministic clusterings, which can be translated in a prior probability distribution defined on the space of all possible deterministic clusterings; in the associated data analysis then an optimal pair of a deterministic clustering and a parameter vector pertaining to the within-cluster distributions is looked at, optimality being defined in terms of a maximum likelihood, a maximum posterior probability or a minimum Bayesian risk or expected loss.

3b) Random-partition clustering methods involve stochastic prior assumptions on a latent crisp cluster membership matrix as well as parametric within-cluster distributional specifications; marginally this can be translated into a mixture distribution. In the associated data analysis, an optimal pair of a parameter vector governing latent cluster membership and of a parameter vector pertaining to the within-cluster distributions is estimated, optimality being defined in terms of a maximum likelihood or a maximum posterior probability criterion; from the resulting parameter estimates one may subsequently derive posterior cluster membership probabilities.

3 Structured overview of two-mode clustering methods

In this section we will present a structured overview or taxonomy of the two-mode clustering area. A schematic overview of this taxonomy (including the corresponding numbers of sections in which the different cells of the taxonomy will be proposed) is given in Table 2. As may be derived from this table, the taxonomy is built on three structuring principles:

1) The primary structuring principle pertains to the *set-theoretical nature of the row and column clusterings* as implied by the different two-mode clustering approaches. Given that for all available two-mode clustering methods the implied row and column clusterings are of the same type (although other possibilities could be considered; see e.g., Figure 1), three families of methods will be distinguished, implying row/column partitionings, nested row/column clusterings and overlapping row/column clusterings, respectively. One may note that this division parallels, to some extent (although not fully), a standard[19,21] three-group classification of the two-mode clustering domain.
2) Within the primary classification, a secondary classification principle is the *set-theoretical nature of the data clusters*. The first and third families of methods are homogeneous in this regard, yielding data cell partitionings and data cell overlapping clusterings, respectively; within the second group of methods, approaches yielding data cell partitionings and two types of nested data cell clusterings may be distinguished.
3) The third classification principle, which will be used to differentiate within the groups obtained from the two previous principles, is the *level of modeling and optimization*.

3.1 Methods that imply row/column partitionings

All methods included in this section imply a partition $(A_1, \ldots, A_r, \ldots, A_R)$ of $\mathcal{R}$, a partition $(B_1, \ldots, B_c, \ldots, B_C)$ of $\mathcal{C}$ and a data clustering that is a partition of $\mathcal{R} \times \mathcal{C}$ as obtained by fully crossing the row and column partionings; the last one means that the data clustering comprises $D = R \times C$ data clusters (blocks) $A_r \times B_c$. A schematic representation of this is presented in Figure 2. The display in this figure presupposes that the rows and columns of the data matrix have been permuted such that all row and column clusters consist of neighboring elements (which can always be done, without loss of generality).

Within the family of row/column partitionings we can draw a further distinction between procedural, deterministic and stochastic approaches. We will now successively deal with each of them.

3.1.1 Procedural methods

From an historical point of view, it is important to include here a reference to work by Lambert and Williams[29,30] on a method they called *nodal analysis*. Stemming from the context of plant ecology, this method aims at retrieving the structure in two-way two-mode plant (species) by vegetation (habitat) presence/absence data. In particular, it aims at constructing two crossed partitions of plants and vegetations, yielding

Table 2 Overview of the proposed taxonomy of two-mode clustering methods with between parentheses the numbers of the sections in which each of the cells of the taxonomy will be described

Classification principle										
1. Set-theoretical nature of row/column clustering	Partitioning (3.1)			Nested clustering (3.2)					Overlapping clustering (3.3)	
2. Set-theoretical nature of data clustering	Partitioning			Partitioning (3.2.1)		Nested clustering (unrestricted) (3.2.2)	Nested clustering (restricted) (3.2.3)		Overlapping clustering	
3. Level of modeling and optimization[a]	1 (3.1.1)	2 (3.1.2)	3 (3.1.3)	1 (3.2.1.1)	2 (3.2.1.2)	1	1 (3.2.3.1)	2 (3.2.3.2)	1 (3.3.1)	2 (3.3.2)

[a] 1 = procedural;
2 = deterministic;
3 = stochastic.

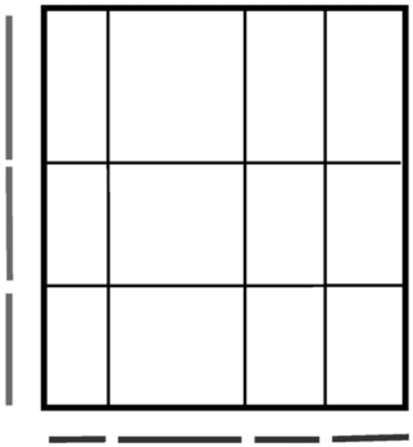

Figure 2 Schematic representation of an hypothetical example of a two-mode partitioning.

a partition into data clusters that are intended to be as homogeneous as possible. Key data clusters are defined as those that contain as many ones as possible and that are called nodes and subnodes (noda/subnoda). The method further particularly aims at retrieving plants and vegetations that are necessarily present in a data cluster [i.e., that correspond to a homogeneous row (respectively, column) pattern of ones within the data cluster]; to achieve this goal, if needed, within data clusters rows or columns that prevent the occurrence of necessary plants/vegetations are left aside. To be sure, the algorithm associated with nodal analysis is of a sequential rather than of a simultaneous type (to the regret of its authors), but the major concern underlying the approach is clearly two-modal (as also appears from the authors' emphasis on the importance of vegetation 'as composed of species-*cum*-habitat units, doubly defined and doubly abstracted from the information in a given set of data'[29])

All further methods to be discussed in the present and the next section aim at approximating the $I \times J$ data matrix $\mathbf{X} = (x_{ij})$ by some $I \times J$ ('reconstructed data') matrix $\hat{\mathbf{X}}$. The entries $\hat{x}_{ij}$ of $\hat{\mathbf{X}}$ are further assumed to be constant within each data cluster (or 'block') $\mathrm{A}_r \times \mathrm{B}_c$ (which implies that the methods under study at least implicitly assume matrix-conditionality of the data). The constancy implies that all rows in a row cluster (and similarly all columns in a column cluster) show identical behavior in $\hat{\mathbf{X}}$. It further follows that the values of the approximating matrix $\hat{\mathbf{X}}$ can be summarized in an $R \times C$ matrix $\mathbf{W}$ of block constants, with:

$$\hat{\mathbf{X}} = \mathbf{AWB}' \tag{1}$$

that is,

$$\hat{x}_{ij} = \sum_{r=1}^{R} \sum_{c=1}^{C} a_{ir} b_{jc} w_{rc}$$

where **A** and **B** are the membership matrices of the row and column partitions and $'$ denotes transpose.

Partitioning methods based on Equation (1) have been extensively studied within the domain of sociometrics under the name of *blockmodeling*.[31–33] Within the blockmodeling approach, the matrix **W** is typically called the image matrix. Researchers of blockmodeling have spent much attention to the problem of reordering (permuting) the rows and columns of a data matrix **X** such that the values in the modified (permuted) data matrix suggest some easily interpretable 'block structure' (e.g., homogeneous blocks with 'high' or 'low' values, 'high values' along the diagonal of **X** – seriation – and so on). To find optimal permutations, several algorithms have been investigated, such as the *bond energy algorithm*.[34–36] The latter algorithm looks across all $I! \times J!$ pairs of row and column permutations for the pair that is such that for the permuted labels i and j the following criterion is optimized:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}[x_{i\,j-1} + x_{i\,j+1} + x_{i-1\,j} + x_{i+1\,j}]$$

with the convention that $0 = x_{0\,j} = x_{i\,0} = x_{I+1j} = x_{iJ+1}$. This criterion intends to place large data values x_{ij} of the data matrix **X** into contiguous cells, eventually near the diagonal of **X**. Apart from the fact that it has been suggested that the bond energy algorithm performs best with binary data, it should be noted that in order to derive exact row and column partitions from the output of this algorithm (i.e., the permuted data matrix), the output has to be subjected to an additional partitioning method.[34] Note that this implies that, from a two-mode clustering perspective, the bond energy approach basically does not involve a global loss or objective function.

3.1.2 Deterministic methods

The blockmodeling (and especially also the bond energy algorithm) approach is closely related to the problem of seriation (i.e., the reconstruction of a lost time order) in archeology. Marcotorchino[37] reviewed a number of techniques in this regard. In this review, he especially focused on seriation methods for binary data on the basis of Equation (1), with **W** being an identity matrix. The latter assumption implies that, upon an appropriate permutation of rows and columns, the reconstructed data take the form of a block diagonal matrix. Marcotorchino discusses algorithms to look for the optimally approximating block diagonal matrix $\hat{\mathbf{X}}$ for a given binary data matrix **X**, optimality being defined in terms of the following least squares loss function:

$$L = \mathrm{trace}[(\mathbf{X} - \hat{\mathbf{X}})(\mathbf{X} - \hat{\mathbf{X}})'] = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \hat{x}_{ij})^2 \qquad (2)$$

Note that if $\hat{\mathbf{X}}$ is defined by Equation (1), then Equation (2) can further be written as:

$$L = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{i \in A_r} \sum_{j \in B_c} (x_{ij} - w_{rc})^2$$

Note also that, in the present binary case, Equation (2) can also be considered a least absolute deviations loss function.

For *binary data*, a simultaneous clustering approach based on Equation (1) and with an unconstrained, possibly rectangular, binary matrix $\mathbf{W}$ has been proposed by Govaert.[38–40] Govaert further proposed an associated algorithm of an iterative dynamic clustering type that optimizes loss function (2).

Also for *real-valued case by variable data*, one may consider clustering methods on the basis of Equation (1), now with an arbitrary, possibly rectangular, real-valued matrix $\mathbf{W}$ and with associated algorithms that optimize Equation (2). For this problem, DeSarbo[41] proposed an alternating least squares algorithm, Gaul and Schader[42] an alternating exchanges algorithm, Baier *et al.*[43] and Vichi[44] a k-means algorithm, Trejos and Castillo[45,46] some algorithms based on taboo search and simulated annealing and Hansohm[47] some genetic algorithms. From his part, Govaert[40] slightly generalized the problem as outlined earlier by introducing weights $\mu_i > 0$ for the cases (i.e., the rows), with $\sum_{i=1}^{I} \mu_i = 1$, and weights $v_j > 0$ for the variables. The core concept behind the loss function he proposes is that of inertia; if the data matrix is column-centered (if necessary after a suitable transformation), that is, if $\forall j$: $\sum_{i=1}^{I} \mu_i x_{ij} = 0$, then the inertia of the data is defined as:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_i v_j x_{ij}^2 \tag{3}$$

The idea further is that a partitioning of cases and variables (as formalized by membership matrices $\mathbf{A}$ and $\mathbf{B}$) implies a reduction of $\mathbf{X}$, the reduction being in terms of a so called summary matrix $\mathbf{W}$, with

$$w_{rc} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} a_{ir} b_{jc} \mu_i v_j x_{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{J} a_{ir} b_{jc} \mu_i v_j}$$

and in terms of case cluster weights $\mu_r^* = \sum_{i=1}^{I} a_{ir} \mu_i$ and variable cluster weights $v_c^* = \sum_{j=1}^{J} b_{jc} v_j$. The inertia of the reduced data then reads as follows:

$$\sum_{r=1}^{R} \sum_{c=1}^{C} \mu_r^* v_c^* w_{rc}^2 \tag{4}$$

The optimal reduction should be such that the loss in inertia implied by the reduction, as formalized by the difference between Equations (3) and (4), is minimal, or, equivalently, such that Equation (4) is maximal. Govaert[40] proposes an iterative dynamic clustering algorithm to achieve this goal. It is interesting to note that the criterion to be optimized by this method is very similar to the one used in principal components analysis and can be shown to be a two-way generalization of the classical *k*-means criterion. Moreover, one may also note that, in case of equal case and variable weights μ_i and v_j, the minimization of the difference between Equations (3) and (4) is equivalent to the minimization of Equation (2), with $\hat{\mathbf{X}}$ being defined in terms of Equation (1) with an unconstrained real-valued matrix **W**.

A final group of cases to be considered pertains to *data of the categorical predictor type*. Within this group, two subgroups of methods can be discerned. First, Bock[48] has proposed an approach on the basis of the *concept of interaction* between the row clusters A_r and the column clusters B_c as measured by:

$$\gamma_{rc} = \bar{x}_{\mathrm{A}_r \times \mathrm{B}_c} - \bar{x}_{\mathrm{A}_r} - \bar{x}_{\mathrm{B}_c} + \bar{x}$$

a formula well known from analysis of variance. Bock then looks for the partitions $(\mathrm{A}_1, \ldots, \mathrm{A}_r, \ldots, \mathrm{A}_R)$ and $(\mathrm{B}_1, \ldots, \mathrm{B}_c, \ldots, \mathrm{B}_C)$ that are such that the following overall interaction criterion

$$\sum_{r=1}^{R} \sum_{c=1}^{C} \#A_r \; \#B_c \gamma_{rc}^2$$

is maximized. In the special case of contingency table data, his approach may be useful for detecting row and column classes that are mutually maximally associated, and hence such that the membership of an object in a row class allows an optimal prediction of its column class, and vice versa.

A second group of approaches pertains to the special case of contingency table data only.[38,40] As in the earlier case, the goal behind the approaches is to *maximize the association* between the row and column classes, strength of association now being captured by the classical χ^2 measure, with:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(x_{ij} - x_{i\cdot} x_{\cdot j}\right)^2}{x_{i\cdot}\, x_{\cdot j}} \tag{5}$$

with $x_{i\cdot}$ (respectively, $x_{\cdot j}$) denoting the marginal row (respectively, column) frequencies. Similarly, we may consider the block frequencies w_{rc} for all data clusters $\mathrm{A}_r \times \mathrm{B}_c$ implied by the partition matrices **A** and **B**, with:

$$w_{rc} = \sum_{i \in \mathrm{A}_r} \sum_{j \in \mathrm{B}_c} x_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} a_{ir}\, b_{jc}\, x_{ij}$$

and with associated χ^2 value:

$$\chi^2 = \sum_{r=1}^{R}\sum_{c=1}^{C}\frac{(w_{rc} - w_{r.}w_{.c})^2}{w_{r.}w_{.c}} \tag{6}$$

The optimal partitionings sought for then are those that minimize the difference between Equations (5) and (6), or, equivalently, that maximize Equation (6). Greenacre[49] proposed a heuristic method to achieve this goal. This method is based on an agglomerative hierarchical clustering of both the rows and the columns of the data matrix **X**, the agglomerative procedure being based on a χ^2 related distance function between row (respectively, column) vectors. Making use of the two hierarchical clusterings, a final split that maximizes Equation (6) is derived. Govaert[38,40] proposed an iterative dynamic clustering procedure to maximize Equation (6). From his part, Bock[50–52] suggested to solve the problem by means of a *maximum-tangent-plane* algorithm. As a matter of fact, Bock has shown that loss function (6) is a member of a broad family of loss functions that measure the discrepancy between the actual aggregated frequency distribution obtained from **X** and the corresponding frequency distribution in the case of independent row and column clusters. These criteria can all be written in a form that involves a *convex function of the partition class centroids* or class frequencies, and maximization then can be achieved through a generic, iterative and geometry-based algorithmic approach. Bock's family of convexity-related two-way clustering criteria also includes instances that formalize a maximum dependency information between classes by using various measures other than the χ^2 criterion (5), such as, for instance, the Kullback–Leibler information measure. Moreover, the family also includes clustering criteria of quite a different nature, such as criteria that attain maximal asymptotic efficiency of a χ^2 test when testing some null distribution against a specific alternative.[53,54] Note that, while distributional assumptions are involved in this last case, the simultaneous classification itself is not part of the stochastic modeling. Note also that still other maximum dependency measures have been studied outside Bock's family (e.g., a variant of Spearman's ρ[55]).

3.1.3 Stochastic methods

Three methods will be discussed in this section. The first one is based on a *random-partition* or mixture modeling approach, the second one on a *fixed-partition* approach and the last one on a *combination of fixed- and random-partition* models.

Within the context of *random partitioning*, DeSarbo *et al.*[56] propose, for data that take values in the interval [0,1], a mixture modeling counterpart of the model based on Equation (1) with a diagonal matrix **W**. The model implies a partition $(A_1, \ldots, A_r, \ldots, A_R)$ of the rows and a partition $(B_1, \ldots, B_c, \ldots, B_C)$ of the columns with $R = C$, and with the assumption that for both modes the unobserved partition class membership data are independently and indentically multinomially distributed. Furthermore, for each row class A_r, two beta distributions are defined for the data x_{ij}, one pertaining to the associated diagonal block data cluster $A_r \times B_r$ and one to the associated off-diagonal data clusters $A_r \times B_c$ with $r \neq c$. DeSarbo *et al.* present a hierarchical Bayesian procedure to estimate this model.

With regard to *fixed partitioning*, Govaert and Nadif[57] proposed stochastic extensions of a model on the basis of Equation (1) with an arbitrary (possibly rectangular) matrix **W**. To be more specific, the authors assume latent partitions **A** and **B** for rows and columns, and suppose that all observations x_{ij} from a resulting data cluster $\mathrm{A}_r \times \mathrm{B}_c$ are independent and identically distributed with a distribution $P(\theta_{rc})$ that includes parameters θ_{rc} that are specific for the data cluster in question. Govaert and Nadif initially discuss both discrete and continuous within-data cluster distributions and both fixed-partition (classification likelihood) and random-partition (mixture modeling) estimating approaches. They then elaborate on the fixed-partition estimation of a model with Bernoulli-distributed data entries, making use of a CEM (classification EM) algorithm.

On the level of *combined fixed- and random-partition* models, Hartigan[58] studied YEA/NAY votes of senators (rows) on various legislative measures (columns). In doing so, he looked for a simultaneous partitioning of rows and columns, on the basis of a fixed-partition approach, in conjunction with a mixture model for the resulting data clusters. In particular, Hartigan assumes a prior probability distribution on the sets of all possible row (respectively, column) partitionings (which is a member of a broader class of so called product partition models[59]). Furthermore, he assumes a two-class random partitioning of the data clusters. For each of the two classes of this random partition a distinct binomial distribution is assumed for the numbers of YEAs and NAYs in the respective data clusters. This implies that for the whole of all data clusters a mixture of two binomials is assumed (which, hopefully, will turn out to be one with a high and one with a low YEA-probability). Hartigan proposed an alternating randomized combinations heuristic to retrieve the partitions and mixture model parameter vector with the highest overall posterior probability.

3.2 Methods that imply nested row/column clusterings

All methods in this section imply a nested row clustering **A** and a nested column clustering **B**. Within this group of methods, three subgroups can be distinguished in terms of which form the *data clustering* takes. A schematic representation of the type of data clustering associated with each of the three subgroups can be seen in Figure 3. In the first subgroup, the data clustering takes the form of a *partition* of $\mathcal{R} \times \mathcal{C}$. In the second and third subgroups, it takes the form of a *nested clustering* of $\mathcal{R} \times \mathcal{C}$, yielding what Hartigan[17] called a '*three-tree*' structure. The difference between the second and third subgroups pertains to an additional restriction with regard to nestedness that is implied by the third subgroup and not by the second. We will now successively discuss each of the three subgroups in more detail.

3.2.1 Data partitionings

This subgroup includes methods on both the procedural and deterministic level of optimization. We will discuss both types of methods successively.

3.2.1.1 Procedural methods

The two methods in this section are called by Hartigan[2] *direct splitting* methods. They do not involve an approximating matrix $\hat{\mathbf{X}}$ for the data matrix **X**. Rather, they aim at data clusters that are such that, within each data cluster E, the variance within each

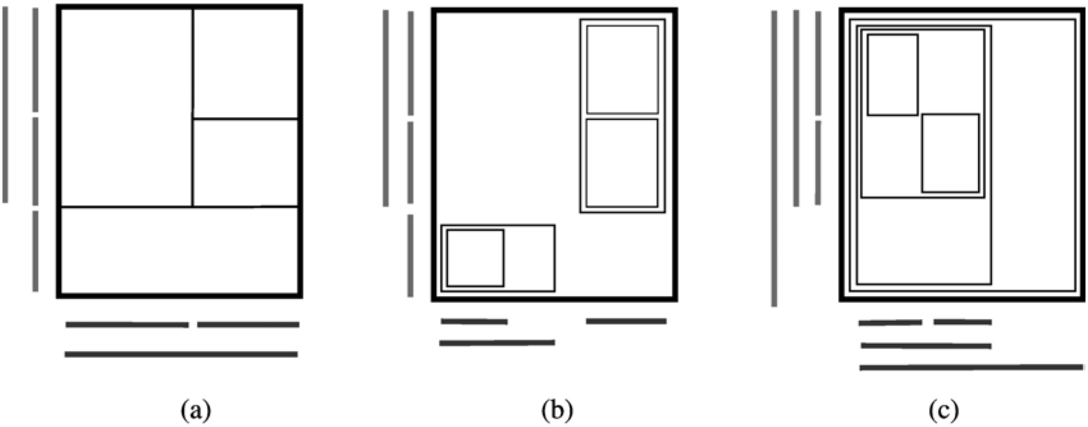

Figure 3 Schematic representation of hypothetical examples of three types of two-mode clustering that imply nested row and column clusterings: (a) data partitioning, (b) unrestricted nested data clustering and (c) restricted nested data clustering.

column (and optionally also within each row) falls below a user-prespecified threshold. In both methods this is achieved through a stepwise splitting procedure. More specifically, the first method, called *one-way splitting*, only aims at low within-column variances; as such, this method presupposes column-conditional data only. On the other hand, the second method, called *two-way splitting*, aims at both low within-column and within-row variances, and therefore requires matrix-conditional data.

3.2.1.2 Deterministic methods

Two more splitting methods have been proposed on a deterministic level. A first method[17] aims at a data partitioning with implied nested row and column clusterings, and with a single approximating data value $\hat{x}_E$ within each data cluster E. Note that the latter presupposes matrix-conditionality of the data. The objective function to be optimized by the associated algorithm is the least squares loss function (2). Hartigan proposed a stepwise splitting procedure for the optimization, with constraints on the successive splits that are such that the nestedness of the implied row and column clusterings is guaranteed throughout the algorithmic process. From their part, Duffy and Quiroz[60] proposed to address the same problem with an alternative successive splitting algorithm that makes use of a permutation distribution to decide where to split and when to stop.

3.2.2 Unrestricted nested data clustering

Within this subgroup, a single method may be situated, as proposed by Hartigan,[2] and called *two-way joining*. This method is to be situated on the procedural level of modeling and optimization. It is an agglomerative procedure that ultimately aims at a minimal number of nested data clusters that are such that any data cluster meets some criterion of homogeneity for its corresponding data values, leaving aside the entries of data clusters that are a subset of it. As a matter of fact, Hartigan developed two variants of this method. A first variant is suitable for (matrix-conditional) binary data; in that case the homogeneity criterion is identical data values. A second variant is suitable for column-conditional continuous data, and aims at data clusters for which the range of data values falls below some prespecified threshold, upon an appropriate rescaling of the columns (variables) such as to make them comparable across rows (cases).

3.2.3 Restricted nested data clustering

All methods reviewed in this section require matrix-conditional data that can be interpreted as proximities. The key common feature of the methods further reads that they all imply a *hierarchical clustering of* $\mathcal{R}\dot{\cup}\mathcal{C}$, that is, a hierarchical clustering of the disjoint union of the row and column modes. Any cluster E of this clustering $\mathcal{H}$ can be written as $E = A_r\dot{\cup}B_c$ with $A_r \subseteq \mathcal{R}$ and $B_c \subseteq \mathcal{C}$, and, hence, implies a row cluster A_r and a column cluster B_c (provided that the latter are non-empty). Furthermore, if both $A_r \neq \emptyset$ and $B_c \neq \emptyset$, then E also implies a data cluster $A_r \times B_c$. It should be clear that the row clustering, column clustering and data clustering implied by a nested clustering $\mathcal{H}$ of $\mathcal{R}\dot{\cup}\mathcal{C}$ are all three nested clusterings. Moreover, the implied data clustering satisfies an additional nestedness restriction. Indeed, while the general nestedness condition for a data clustering E reads as follows:

$$\begin{aligned}\forall A_r \times B_c, A_{r'} \times B_{c'} \in E:\quad & A_r \times B_c \cap A_{r'} \times B_{c'} \neq \emptyset \\ & \qquad\Rightarrow \\ & A_r \times B_c \subseteq A_{r'} \times B_{c'} \quad \text{or} \quad A_r \times B_c \supseteq A_{r'} \times B_{c'}\end{aligned}$$

in the present case the following stronger condition has to be met:

$$\begin{aligned}\forall A_r \times B_c, A_{r'} \times B_{c'} \in E:\ & A_r \cap A_{r'} \neq \emptyset \quad \text{or} \quad B_c \cap B_{c'} \neq \emptyset \\ & \qquad\qquad\Rightarrow \\ & A_r \times B_c \subseteq A_{r'} \times B_{c'} \quad \text{or} \quad A_r \times B_c \supseteq A_{r'} \times B_{c'}\end{aligned}$$

The latter implies, for example, that it is impossible to have two data clusters $A_r \times B_c$ and $A_{r'} \times B_c$ with $A_r \cap A_{r'} = \emptyset$ (as displayed in the middle panel of Figure 3).

Furthermore, in the hierarchical clustering $\mathcal{H}$ of $\mathcal{R}\dot{\cup}\mathcal{C}$, a hierarchical level function $h\colon \mathcal{H} \to [0, \infty[$ can be defined which is such that:

$$\forall H_1, H_2 \in \mathcal{H}: H_1 \subseteq H_2 \Rightarrow h(H_1) \leq h(H_2)$$

From the hierarchical level function, one may further derive a weight w(H) for each (nonuniversal) data cluster as the absolute difference in hierarchical level between that cluster and the cluster immediately above it in the hierarchy.

As in the two-way one-mode case, a hierarchical clustering and its associated hierarchical level function can be graphically displayed in a tree diagram or dendrogram, the only major difference now being that the leaves are both row and column entries. Also as in two-way one-mode clustering, two cases can further be distinguished:

1) The *ultrametric tree* case for which (in case of dissimilarity data) it holds that:

$$\hat{x}_{ij} = \min_{\substack{\mathrm{H}\in\mathcal{H} \\ (i,j)\in\mathrm{H}}} [h(\mathrm{H})] = h_{\max} - \sum_{\substack{\mathrm{H}\in\mathcal{H} \\ (i,j)\in\mathrm{H}}} w(\mathrm{H}) \tag{7}$$

2) The *additive tree* case for which (in case of dissimilarity data) it holds that:

$$\hat{x}_{ij} = \sum_{\substack{\mathrm{H}=\mathrm{A}_r\times\mathrm{B}_c\in\mathcal{H} \\ (i,j)\in[(\mathrm{A}_r\times\mathcal{C})\Delta(\mathcal{R}\times\mathrm{B}_c)]}} w(\mathrm{H}) \tag{8}$$

where Δ denotes symmetric difference [i.e., $\mathrm{A}\Delta\mathrm{B} = (\mathrm{A}\cup\mathrm{B})\setminus(\mathrm{A}\cap\mathrm{B})$].

The two cases described in Equations (7) and (8), which are rooted in the seminal work of Furnas,[61] are two-mode extensions of the ultrametric and additive tree models for two-way one-mode proximities. Equalities (7) and (8) also imply that for the approximating matrix $\hat{\mathbf{X}}$ two-mode generalizations of the ultrametric inequality and the additive inequality (or four-point condition) apply.[62] Finally, as in the case of two-way one-mode models, if data clusters are interpreted as features that (row/column) entries may or may not possess, Equation (7) may be interpreted as if the approximating proximity between some pair of elements is an additive function of the weights of their common features, whereas Equation (8) means that it is an additive function of their distinctive features (in line with a general feature based account of similarity[63]).

In general, both ultrametric and additive trees suffer from several types of nonuniqueness: First, for both kinds of tree, clusters that contain elements of one mode only are not involved in Equations (7) and (8) and therefore do not influence the approximating matrix $\hat{\mathbf{X}}$; as such, the subtrees consisting of such clusters are arbitrary. Secondly, as in the two-way one-mode case, the placement of the root in additive trees does not influence the end result of Equation (8), and as such is arbitrary.

With regard to algorithms to fit ultrametric and additive trees to two-way two-mode data, the situation is essentially similar to that of two-way one-mode proximities: On the one hand, procedural methods may be distinguished, which do not rely on a global loss function that is implied by the ultrametric or additive tree structure. On the other hand, a few deterministic procedures have been developed that directly optimize a least squares loss function (2) for reconstructed data $\hat{\mathbf{X}}$ on the basis of Equation (7) or Equation (8). Incidentally, while procedural approaches have dealt with the fitting of ultrametric trees, direct optimization methods have addressed the fitting of both ultrametric and additive trees. We will now successively discuss both types of methods briefly.

3.2.3.1 Procedural methods

As stepwise procedural methods, both *simple agglomerative* and *missing value approaches* have been proposed. A simple agglomerative approach has been proposed by Eckes and Orlik,[21,64,65] aiming at data clusters $\mathrm{A}_r \times \mathrm{B}_c$ with minimal internal heterogeneity, heterogeneity being formalized as:

$$\frac{1}{\#(\mathrm{A}_r \times \mathrm{B}_c)} \sum_{i \in \mathrm{A}_r} \sum_{j \in \mathrm{B}_c} (x_{ij} - m)^2 \tag{9}$$

with m denoting the maximum entry in $\mathbf{X}$. This approach has been modified by Mirkin *et al.*[23] so as to turn it into a stepwise procedure to optimize a global loss function on the basis of Equation (9). From their part, Castillo and Trejos[66] generalized the approach of Eckes and Orlik from the point of view of a generalization of the Lance–Williams[67] recurrence formula that underlies many stepwise procedural methods for agglomerative hierarchical clustering of two-way one-mode proximities.

Missing value approaches rely on the fact that two-mode proximity data can be considered incomplete, in that a full set of proximities would pertain to all pairs in $\mathcal{R} \dot{\cup} \mathcal{C}$ (Section 2.1). They either try to complete the missing information prior to a further analysis,[68] or try to estimate it in the course of a stepwise algorithmic procedure.[69]

3.2.3.2 Deterministic methods

All direct optimization procedures are based on the finding that the fitting of ultrametric and additive trees to a given data set $\mathbf{X}$ comes down to a *constrained optimization problem*, the loss function being equal to Equation (2) and the constraints forcing the reconstructed data $\hat{\mathbf{X}}$ to satisfy the two-mode generalization of the ultrametric (respectively, additive) inequality as implied by Equation (7) [respectively, Equation (8)]. Two types of approaches have been proposed to optimize this problem: a *penalty function approach* (described by De Soete and Carroll[62,70]), in which the loss function is augmented with a penalty term that formalizes the constraints, and an *iterative projection approach* (described by Hubert and Arabie[71]), in which tentative solutions are iteratively projected on convex cones that constitute a geometric representation of each of the constraints under study.

3.3 Methods that imply overlapping row/column clusterings

All methods in this section imply an overlapping row clustering, an overlapping column clustering and an overlapping data clustering. A schematic representation of all this is presented in Figure 4. The methods can further be subdivided into: 1) a miscellaneous category that includes a few procedural methods and 2) a main category that contains two families of deterministic approaches. We will now successively discuss each of these two categories of methods.

3.3.1 Procedural methods

From a historical point of view, it is important to make a reference to Hartigan's *modal block method*.[72] This method has been developed for column-conditional case by variable data, the variables being categorical. Its underlying principle is to retrieve a small number of data clusters, which are such that within each cluster almost all values for each variable are equal to the modal value within the cluster in question for that variable (excluding possible subclusters). The approach clearly allows for overlap

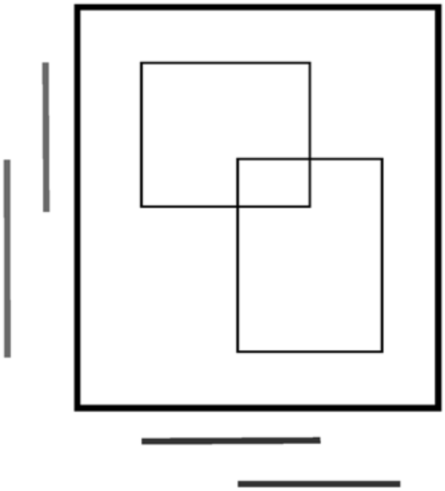

Figure 4 Schematic representation of an hypothetical example of a two-mode overlapping clustering.

between the data clusters, although the author considers this something like a nuisance, as he thinks this may hamper the interpretation. More recently, Eckes and Orlik[21] and Mirkin *et al.*[23] have proposed overlapping clustering variants of their procedural two-mode hierarchical clustering methods as discussed in Section 3.2.3.

3.3.2 Deterministic methods

All methods in this section imply a matrix $\hat{\mathbf{X}}$ that is to approximate the data matrix $\mathbf{X}$ as well as possible (in the least squares or the least absolute deviations sense). The entries of $\hat{\mathbf{X}}$ are obtained from a set of constants, a constant being associated with each of the data clusters, in: 1) a *simple additive* or 2) a *Boolean* (or generalized Boolean) way. We will now successively discuss both types of approaches.

3.3.2.1 Additive approaches

For two-way one-mode $I \times I$ similarity data, Shepard and Arabie[73] proposed the *ADCLUS* (additive clustering) model. In matrix form, this model reads as follows:

$$\hat{\mathbf{X}} = \mathbf{AWA}' + \mathbf{C} \tag{10}$$

where $\mathbf{A}$ denotes an $I \times R$ row cluster membership matrix (pertaining to possibly overlapping clusters), $\mathbf{W}$ is a diagonal matrix with positive elements and $\mathbf{C}$ is a matrix with a constant value. If the clusters are interpreted as features and the diagonal values of $\mathbf{W}$ as feature weights, then Equation (10) simply means that the similarity between two objects equals (upon an additive constant) the summed weights of their common features.

For two-way two-mode data, various generalizations of the ADCLUS model have been advanced. All of them assume that the data can be interpreted as proximities, and are based on the following straightforward generalization of Equation (10):

$$\hat{\mathbf{X}} = \mathbf{A}\mathbf{W}\mathbf{B}' + \mathbf{C} \tag{11}$$

In Equation (11), **A** and **B** denote overlapping row and column clusterings; moreover, the non-zero elements w_{rc} correspond to data clusters $\mathrm{A}_r \times \mathrm{B}_c$; Equation (11) therefore means that, for a row i and a column j, $\hat{x}_{ij}$ equals the sum of the weights of all data clusters to which (i, j) belongs. In the special case that **W** is diagonal (which implies a 1–1 relation between row and column clusters), the row and column clusters can be interpreted as corresponding features; Equation (11) then again means that $\hat{x}_{ij}$ equals the summed weights of the common features of i and j; the latter generalizes the two-mode ultrametric tree model by removing the nestedness restrictions on **A**, **B** and the data clusters.

With regard to methods based on Equation (11), Mirkin *et al.*[23] proposed *additive box clustering*, which implies Equation (11) with **W** constrained to be diagonal. It may be noted that the same model can also be obtained as a special case of *CANDCLUS* (CANonical Decomposition CLUStering), a general clustering and multidimensional scaling approach for multiway data as proposed by Carroll and Chaturvedi.[74] Furthermore, DeSarbo[41] introduced *GENNCLUS* (GENeral Nonhierarchical CLUStering). In this method, **W** is constrained to be square and symmetric (but not necessarily diagonal). GENNCLUS includes possibilities for various kinds of constraints, including the constraint that **A** and **B** are partition matrices (Section 3.1.1) and the constraint that **W** is diagonal. Finally, Gaul and Schader[42,43] introduced *PENCLUS* (PENalty CLUStering). This approach does not imply any restrictions on **W** in Equation (11), but leaves it as an option to constrain **A** and **B** to partition matrices (Section 3.1.1). All of the approaches as mentioned earlier are associated with iterative alternating algorithms that optimize the least squares loss function (2).

3.3.2.2 (Generalized) Boolean approaches

This last group of methods roots in the modeling of binary two-way two-mode data, although, as will be mentioned subsequently, generalizations to rating and ordinal data are now available as well. A somewhat older method within this group is *Boolean factor analysis*.[75] The latter can again be described as in Equation (11), with **W** now being an $R \times R$ identity matrix, and, more importantly, now a Boolean rather than an ordinary matrix product (which we will further denote by $\otimes$):

$$\hat{\mathbf{X}} = \mathbf{A} \otimes \mathbf{B}' \tag{12}$$

Note that a Boolean matrix product is calculated as an ordinary matrix product, but with a Boolean sum $\oplus$, which implies that $1 \oplus 1 = 1$. Once again, the columns of **A** and **B** denote membership in (possibly overlapping) row and column clusters; alternatively, they can also be interpreted as binary factor loadings and scores. The data clusters further are the R sets $\mathrm{A}_r \times \mathrm{B}_r$. If we denote the column vectors of **A** by

$\mathbf{a}_{.1}, \ldots, \mathbf{a}_{.r}, \ldots, \mathbf{a}_{.R}$ and those of $\mathbf{B}$ by $\mathbf{b}_{.1}, \ldots, \mathbf{b}_{.r}, \ldots, \mathbf{b}_{.R}$, then Equation (12) can be rewritten as:

$$\hat{\mathbf{X}} = \bigoplus_{r=1}^{R} \mathbf{a}_{.r} \otimes \mathbf{b}'_{.r} \tag{13}$$

which means that the data cells with one-entries in $\hat{\mathbf{X}}$ are the union of R rectangles (i.e., the R data clusters). Mickey *et al.*[75] developed an iterative algorithm to fit Equation (12) to a given binary matrix with a least squares (or, equivalently, a least absolute deviations) loss function (2).

De Boeck and Rosenberg[76] modified the Boolean factor model in an important respect. As such, these authors had a special interest in two types of relations among the row and column entities of $\hat{\mathbf{X}}$: an equivalence relation, equivalence meaning that the corresponding rows (columns) are identical, and an if–then type implication relation, implication meaning that the pattern of ones for one row (column) is a subset of the pattern of ones for another row (column). De Boeck and Rosenberg wanted the row vectors of **A** and **B** to reflect these two types of relations. The latter means that equivalent entities in $\hat{\mathbf{X}}$ should have identical row vectors in **A** (**B**) and that an implication relation between two entities should be reflected in a subset–superset relation between their row vectors in **A** (**B**). All this implies that the partitions that may be derived from **A** and **B** (section 2.1) necessarily represent the equivalence classes of the two equivalence relations, and, moreover, that the two partitions are hierarchically organized in terms of the if–then type implication relations as mentioned earlier (for a discussion of the use of the concept of hierarchy here refer to Van Mechelen *et al.*[77]). De Boeck and Rosenberg called the resulting model a *hierarchical classes model*, and devised a comprehensive graphical representation of it that includes both a representation of the hierarchically organized partitions of row and column entries and a representation of the linkage between the row and column hierarchies as implied by Equation (12). On an algorithmic level, De Boeck and Rosenberg developed a procedure to fit hierarchical classes models to a given data set, making use of loss function (2); this procedure consists of an improved version of the Boolean factor analysis algorithm, followed by an additional routine to guarantee the correct representation of the equivalence and implication relations in **A** and **B**. Interestingly, by the latter routine the rectangles implied by Equation (13) are turned into maximal rectangles within $\hat{\mathbf{X}}$ (without affecting the $\hat{\mathbf{X}}$-matrix resulting from the improved Boolean factor analysis routine).

Since the paper by De Boeck and Rosenberg, the hierarchical classes approach has been extended in many ways. As such, a dual version of the original model has been developed, called the conjunctive hierarchical classes model,[78] which, amongst other things, allows the derivation of implications between conjunctions of entities (e.g., *if* variables 1 *and* 2 are present, *then* variable 3 should also be present). (For more recent algorithmic work refer to Leenen and Van Mechelen.[79]) An important recent extension of the approach is for matrix-conditional rating-valued data; this extension includes, in addition to the overlapping clustering matrices **A** and **B**, a (possibly rectangular) rating-valued 'core' matrix **W** (data not published).

From a theoretical point of view, it is important to refer to the connection between the hierarchical classes model on the one hand, and the set of all maximal one-rectangles within a binary data set on the other hand. On the set of maximal rectangles, a partial order relation can be defined in terms of subset–superset relations between the row (or column) clusters implied by the maximal rectangles; in this way the set is turned into a lattice. This kind of lattice has been extensively studied under the names of *Galois lattice* (within a French research tradition[80]) and of *formal concept lattice* (within a German research tradition[81]). Also extensions for ordinal data have been developed.[82] From a logical point of view, it is interesting to note that the Galois lattice of formal concepts represents the full set of implications among conjunctions of row (column) entities. A point of difference between the hierarchical classes and lattice approaches is that the latter do not include a decomposition of the binary data set under study as in Equation (12) (although such a decomposition could be added to the lattice framework[83]). The most important difference between the two approaches, however, is that the lattice approach is restricted to exact representations of the data (i.e., $\mathbf{X} = \hat{\mathbf{X}}$). Apart from the fact that such a restriction may be less appropriate in case of non-errorfree data, this further also implies that, except in the case of very small data sets, the resulting lattices will typically be very complicated and, hence, hard to interpret. One possible way out in this regard could be to subject the data first to a hierarchical classes analysis (with a low number of overlapping clusters R), and to subsequently construct the lattice of the reconstructed data matrix $\hat{\mathbf{X}}$ as implied by the hierarchical classes model.[84]

4 Illustrative applications

The data set we will use in this section was taken from work by Mezzich and Solomon.[85] Each of 22 experienced psychiatrists was invited to think of a typical patient for each of four diagnostic categories: manic-depressive depressed (MDM), manic-depressive manic (MDD), simple schizophrenic (SS) and paranoid schizophrenic (PS); subsequently each patient had to be characterized by 0–6 severity ratings on 17 psychiatric symptoms. The sympoms were: somatic concern, anxiety, emotional withdrawal, conceptual disorganization, guilt feelings, tension, mannerisms and posturing, grandiosity, depressive mood, hostility, suspiciousness, hallucinatory behavior, motor retardation, uncooperativeness, unusual thought content, blunted affect, excitement. (For a more extensive study within the same domain, making use of a two mode clustering procedure but based on an official psychiatric diagnostic manual rather than on clinician judgments, refer to Gara *et al.*[86]). To be sure, the data of Mezzich and Solomon could be considered three-way three-mode, $x\colon \mathrm{S}_1 \times \mathrm{S}_2 \times \mathrm{S}_3 \rightarrow \mathrm{Y}$, with S_1 denoting the set of 22 psychiatrists, S_2 the set of four diagnostic categories, S_3 the set of 17 symptoms and Y the set $\{0,1,2,3,4,5,6\}$. However, in line with the approach taken by Mezzich and Solomon, we will rewrite the data as a two-way two-mode matrix by considering $\mathrm{S}_1 \times \mathrm{S}_2$ as a single set S_1^*. The elements of S_1^*, that is, the rows of the resulting data matrix, are called archetypal patients; in fact they refer to combinations of a psychiatrist and a diagnostic category. For the remainder of this section it is important to note that the set of 88 archetypal patients has a clear underlying four-class partition structure based on the four diagnostic categories under study. Furthermore, in order to study replicability, Mezzich and Solomon randomly divided each class of 22 patients into two groups of 11 each, yielding two data sets of 44 archetypal patients. For the sake of simplicity, we limited our reanalyses to the first data set of Mezzich and Solomon.

We subjected the 44×17 archetypal patient by symptom data set to three series of analyses: 1) one series of two-mode partitionings, 2) one series of two-mode ultrametric tree fittings and 3) one series of rating-valued hierarchical classes analyses. We will now successively discuss each of these analyses more in detail.

As regards *two-mode partitioning*, we will discuss the results with four patient and four symptom clusters. The solution that will be reported is the optimal one out of 100 runs of an alternating exchanges procedure, 100 runs of a tabu search procedure and 100 runs of a simulated annealing procedure, all implemented by Castillo and Trejos.[46] The optimal solution, with a percentage of variance accounted for of 55.2, is displayed in Figure 5. In this and the following two figures, clusters of archetypal patients are simply summarized in terms of frequencies of the four diagnostic categories under study, whereas symptom clusters are displayed enumeratively. The middle part of Figure 5 further shows the matrix **W** with the block constants.

For the *two-mode ultrametric tree-fitting*, the data were first transformed into dissimilarities by subtracting each entry from the maximum possible value, 6. Subsequently, the resulting transformed matrix was subjected to an iterative projection algorithm as implemented by Hubert.[71] The best solution was chosen from 300 runs with a random start. This solution, with a percentage of variance accounted for of 50.7, is partially displayed in Figure 6. The display is partial in that the full two-mode ultrametric tree has 25 leaves; for sake of simplicity Figure 6 shows only the part of the tree between the root and the 11-class partition of $\mathcal{R} \cup \mathcal{C}$. Note that the values of the approximating matrix $\hat{\mathbf{X}}$, that is, the resulting two-mode ultrametric, can immediately be derived from the figure: for example, the ultrametric distance between all SS (in the two upper partition classes) and the symptoms of somatic concern and anxiety is the lowest level at which they are joined in the dendrogram, that is, 3.5; in terms of the original 0–6 symptom severity scale, this comes down to a severity of $6–3.5 = 2.5$.

For the *rating-valued hierarchical classes analysis* we will discuss a solution with three patient clusters and three symptom clusters (as well as two non-zero values in the core matrix **W**). This solution was obtained by means of the HICLAS-R algorithm as implemented by Ceulemans,[87] making use of four sets of rationally derived starting configurations; the solution (which has a Jaccard goodness-of-fit value of 0.673) is displayed in Figure 7. The rectangular boxes in this figure denote the partition classes of patients and symptoms that may be derived from the corresponding overlapping clusterings. Hierarchical (if–then type) relations between the patient partition classes are displayed bottom up and similar relations between the symptom partition classes are displayed top down; for example, from the figure one may derive that the 9 paranoid schizophrenics (PS) in the upper left patient class obtain for each symptom at least the same severity rating (or a higher one) when compared with the 8 simple schizophrenics (SS) in the middle bottom class of the patient hierarchy. The overlapping clusters of patients (symptoms) can further be derived from the figure as each of the bottom classes in a hierarchy together with all of its hierarchically higher classes; for

	11 MDD	10 MDM	11 SS	11 PS 1 MDM
somatic concern **guilt** **depressive mood** **motor retardation**	5.4	0.6	2.1	1.4
tension **grandiosity** **hostility** **excitement**	1.2	5.2	1.1	4.5
emotional withdrawal **tension**	3.6	0.1	5.4	2.4
anxiety **conceptual disorganization** **mannerisms** **suspicious** **hallucinations** **uncooperative** **unusual thought**	2.2	2.3	2.4	4.6

Figure 5 Two-mode partitioning of archetypal psychiatric patient data of Mezzich and Solomon[86] with four patient clusters (displayed in the column headings with MDD, manic-depressive depressed; MDM, manic-depressive manic; SS, simple schizophrenic and PS, paranoid schizophrenic) and four symptom clusters (displayed in the row headings) and matrix **W** of block constants (displayed in the body of the figure).

example, the first overlapping patient cluster comprises all manic-depressive manic (MDM) patients as well as all PS. Finally, the values of the approximating matrix $\hat{\mathbf{X}}$ may be derived from the figure as follows: The association strength $\hat{x}_{ij}$ of a given patient i and a given symptom j corresponds to the highest value on a path linking them; for example, the 9 PS in the upper left patient class of the patient hierarchy are associated with severity '5' with grandiosity, conceptual disorganization, anxiety and so on, and with severity '2' with mannerisms, somatic concern and so on.

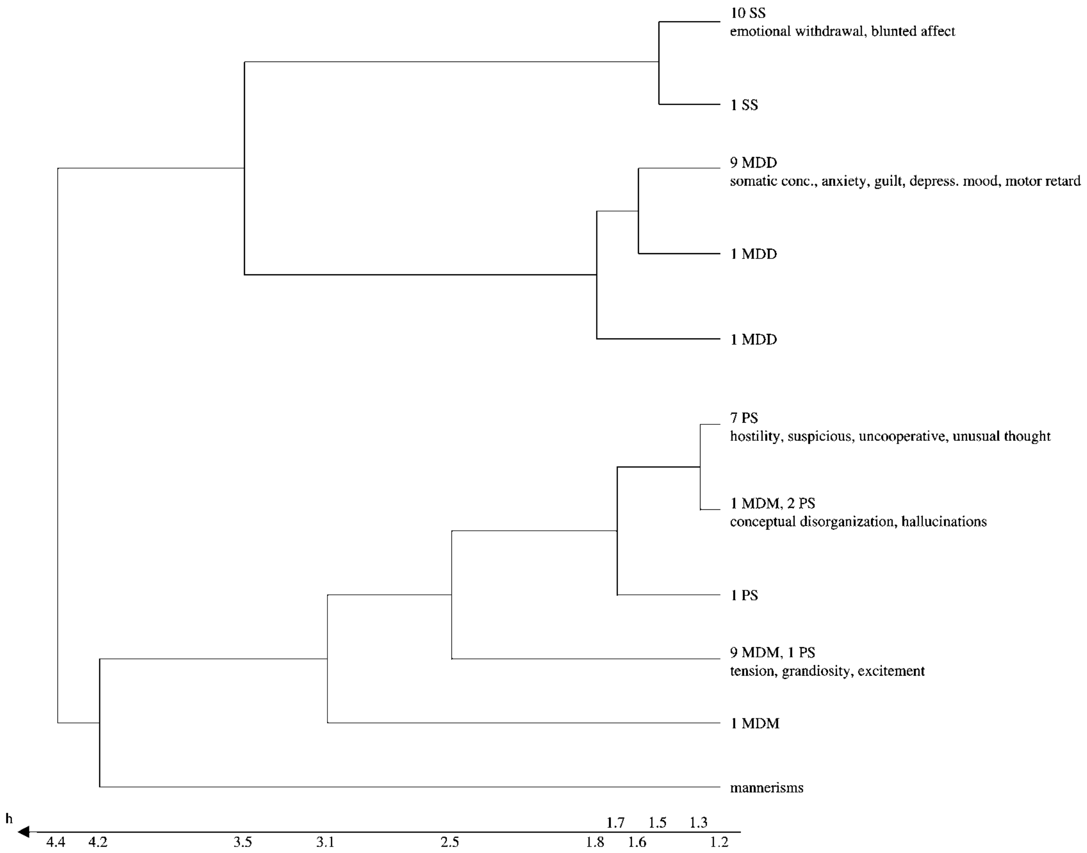
10 SS
emotional withdrawal, blunted affect
1 SS
9 MDD
somatic conc., anxiety, guilt, depress. mood, motor retard
1 MDD
1 MDD
7 PS
hostility, suspicious, uncooperative, unusual thought
1 MDM, 2 PS
conceptual disorganization, hallucinations
1 PS
9 MDM, 1 PS
tension, grandiosity, excitement
1 MDM
mannerisms
h
1.7
1.5
1.3
4.4
4.2
3.5
3.1
2.5
1.8
1.6
1.2

Figure 6 Two-mode hierarchical clustering of archetypal psychiatric patient data of Mezzich and Solomon[86] (part between root and 11-class partition of $\mathcal{R} \cup \mathcal{C}$), with h denoting hierarchical level and MDD, manic-depressive depressed; MDM, manic-depressive manic; SS, simple schizophrenic and PS, paranoid schizophrenic.

The major results of the three types of analyses can be summarized as follows: All methods yield clinically meaningful clusters and associations. All methods further also perform reasonably well with regard to the recovery of the underlying four-class patient partition structure. With regard to the latter, the performance of the two-mode partitioning methods is clearly superior to the other two methods, with only a single misclassified archetypal patient. (It may be noted that the latter finding is also in line with the results of the initial analyses of the same data by Mezzich and Solomon,[85] who found that a one-mode *k*-means method outperformed various other one-mode clustering methods; the advantage of the present two-way partitioning analysis over the earlier one-mode *k*-means analysis is the clear interpretation of the patient partitioning based on the symptom partitioning.)

The major additional gain that may derived from the output of the two-mode ultrametric tree analysis and rating-valued hierarchical classes analysis pertains to a

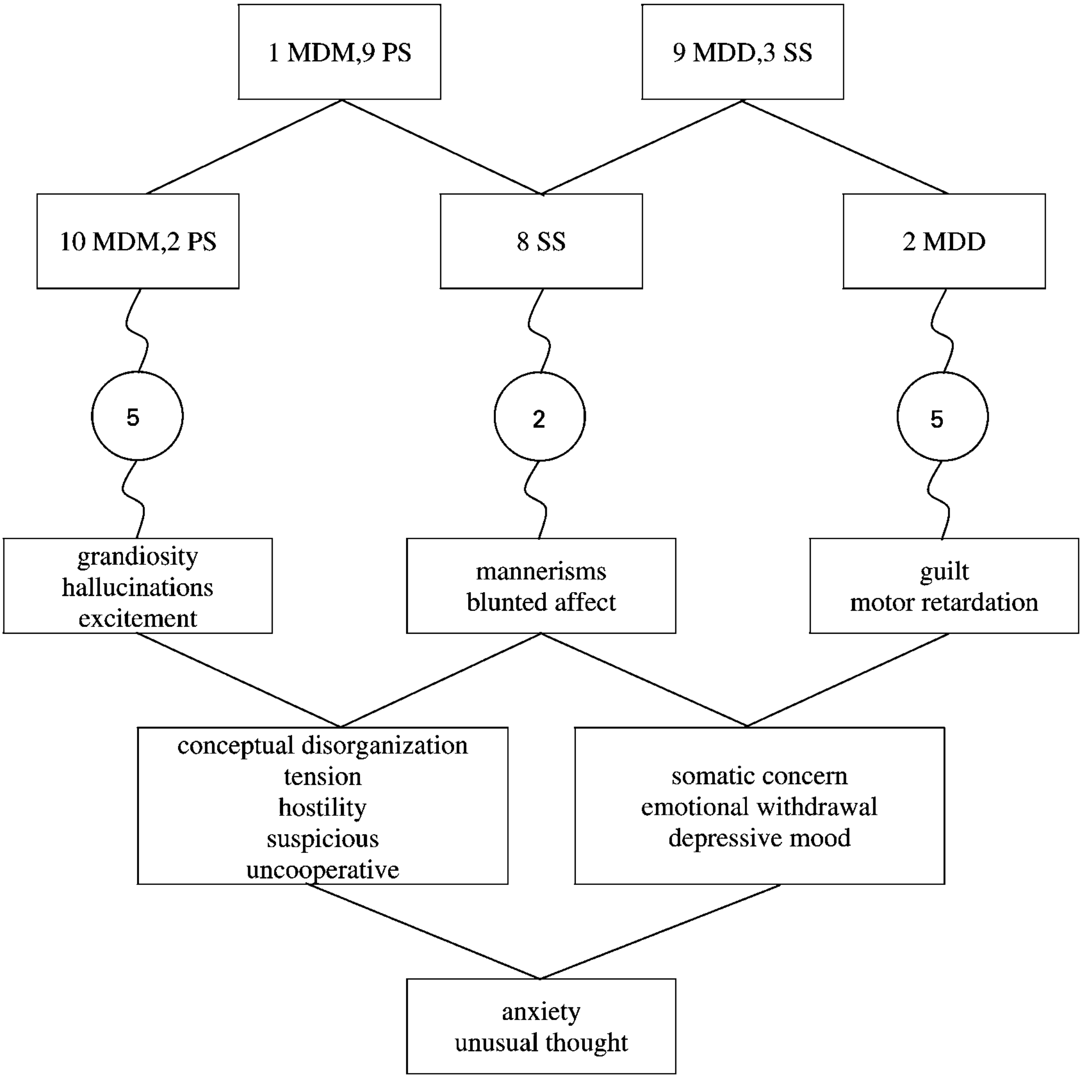

Figure 7 Rating-valued hierarchical classes model for archetypal psychiatric patient data of Mezzich and Solomon[86] with three overlapping clusters of patients (displayed in the upper part of the figure, with MDD, manic-depressive depressed; MDM, manic-depressive manic; SS, simple schizophrenic and PS, paranoid schizophrenic), three overlapping clusters of symptoms (displayed in the lower part of the figure) and values of the core matrix **W** (displayed in the middle part of the figure); boxes indicate the classes of the partitions that may be derived from the overlapping clusterings, hierarchical (if–then type) relations between patient partition classes are to be read bottom up and between symptom partition classes top down.

deeper insight into the structural relations between (sub)groups of patients and between (sub)groups of symptoms; for example, from the HICLAS output one may derive that the diagnostic category of SS largely subsumes the categories of the paranoid schizophrenics and of manic-depressive depressed (MDD) patients.

5 Concluding remarks

On a theoretical level, the variety of two-mode clustering models is large and rich. In spite of a fairly high heterogeneity, it appears from our overview that many models and structures bear interesting relationships. For instance, the key Equations (1), (11) and (12) observed in many models as discussed in this paper are closely related. This may give rise to the conjecture that many of the models under study are even more strongly interrelated on a deeper mathematical level; it would at least be worthwhile to explore the latter issue. Another obvious extension that could be explored in future research is that to the simultaneous clustering of n-way n-mode data with $n \geq 2$.

On a more practical level, the taxonomy as presented earlier also contains a number of useful clues for a concrete data analysis, for instance with respect to the choice of a clustering method in a specific application. First, the nature of the data should be carefully inspected. A key aspect in this regard, as appears from our taxonomy, is the conditionality of the data. Secondly, important choices pertain to the nature of the mathematical structure or model one is willing to consider for the data at hand. It is of course of utmost importance that these choices are in line with the substantive-theoretical concerns and assumptions underlying one's research. Consider, for instance, the choice as to whether the matrix $\mathbf{W}$ in Equation (1) and (11), which constitutes the linking structure between the row and column clustering, has to be a diagonal matrix; the latter comes down to the choice of a one-to-one relationship between row and column clusters. Take, as an example, case by variable data, with cases being patients and variables symptoms. If the theoretical idea underlying the clustering looked for is a set of (latent) syndromes, with patient clusters denoting groups of patients suffering from each of the syndromes, and symptom clusters denoting the groups of symptoms that constitute each of the syndromes, then a one-to-one relation between patient and symptom clusters seems to be called for, indeed, as the concept of a syndrome directly links the two types of clusters. Another choice to be made with regard to the type of mathematical structure underlying the clustering method one may wish to use is that between partitions and overlapping row/column clusterings. If, again in the patient by symptom example, there are good theoretical reasons to allow for syndrome comorbidity and for symptom overlap between syndromes, then the correct choice is to go for overlapping clusterings of both patients and symptoms.

A final practical caveat that should be mentioned is that, whereas general purpose statistical packages include various modules for one-way clustering (albeit usually only on a very elementary level), two-mode methods (implemented as stand-alone programs by their developers) mostly have not been incorporated in these systems. Exceptions include variants of Hartigan's[2] two-way joining (as discussed in Section 3.2.2), which have been incorporated within *Statistica*™ and *Systat*™, and Boolean factor analysis (as discussed in Section 3.3.2.2), which has been incorporated within *BMDP*™. Yet, on a more specialized level, several software packages developed in bioinformatics contain two-mode clustering procedures. Examples include *BioMine 1.0* (option BiCluster), *Coupled Two-Way Clustering, GeneViz 2.0* and *SpotFire Decision Site for Functional Genomics*. This links up with the increased importance of the simultaneous clustering of genes and patients (samples, tissues) in microarray data analysis, as discussed at the beginning of this paper.

Acknowledgement

The research reported in this paper was supported in part by grant IAP P5/24. The authors gratefully acknowledge the assistance of Eva Ceulemans, Javier Trejos and William Castillo in the analyses of the empirical example as reported in the paper. They also wish to thank Larry Hubert for making his software available, and Peter Kuppens and Eva Ceulemans for their comments on a previous version of the manuscript.

References

1 Bock H-H. *Automatische Klassifikation (Clusteranalyse)*. Göttingen: Vandenhoeck & Ruprecht, 1974.

2 Hartigan JA. *Clustering algorithms*. New York: John Wiley, 1975.

3 Arabie P, Hubert LJ, De Soete G. eds. *Clustering and classification*. River Edge, NJ: World Scientific, 1996.

4 Everitt BS, Landau S, Leese M. *Cluster analysis*, fourth edition. New York: Edward Arnold, 2001.

5 Jain AK, Dubes RC. *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

6 Celeux G, Diday E, Govaert G, Lechevallier Y, Ralambondrainy H. *Classification automatique des données: environnement statistique et informatique*. Paris: Dunod, 1989.

7 Kaufman L, Rousseeuw PJ. *Finding groups in data: an introduction to cluster analysis*. New York: Wiley, 1990.

8 Mirkin BG. *Mathematical classification and clustering*. Dordrecht: Kluwer, 1996.

9 Gordon AD. *Classification*, second edition. Boca Raton, FL: Chapman & Hall/CRC, 1999.

10 Tryon RC. *Cluster analysis*. Ann Arbor, MI: Edwards Brothers, 1939.

11 Fisher W. *Clustering and aggregation in economics*. Baltimore: Johns Hopkins, 1969.

12 Getz G, Levine E, Domany, E. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the USA* 2000; **97**(22): 12079–84.

13 Li J, Zha H. Simultaneous classification and feature clustering using discriminant vector quantization with applications to microarray data analysis. *IEEE Computer Society Bioinformatics Conference (CSB'02)* 2002, Stanford, CA.

14 Pollard KS, van der Laan MJ. Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences* 2002; **176**: 99–121.

15 Pollard, KS, van der Laan MJ. Statistical inference for simultaneous clustering of gene expression data. In Denison DD, Hansen MH, Holmes C, Mallick B, Yu B, eds. *Nonlinear estimation and classification*. Berlin: Springer, 2002, 305–320.

16 Jörnsten R, Yu, B. Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics* 2003; **19**: 1100–09.

17 Hartigan J. Direct clustering of a data matrix. *Journal of the American Statistical Association* 1972; **67**: 123–29.

18 Bock H-H. *Stochastische Modelle für die einfache und doppelte Klassifikation von normalverteilten Beobachtungen*. Dissertation, University of Freiburg, Germany, 1968.

19 Both M, Gaul W. Ein vergleich zweimodaler Clusteranalyseverfahren. *Methods of Operations Research* 1987; **57**: 593–605.

20 Eckes T. Bimodale Clusteranalyse: Methoden zur Klassifikation von Elementen zweier Mengen. *Zeitschrift für Experimentelle und Angewandte Psychologie* 1991; **XXXVIII**: 201–25.

21 Eckes T, Orlik P. An error variance approach to two-mode hierarchical clustering. *Journal of Classification* 1993; **10**: 51–74.

22 Krolak-Schwerdt S. Two-mode clustering methods: compare and contrast. In Schader M, Gaul W, Vichi M, eds. *Between data science and applied data analysis: studies in classification, data analysis, and knowledge organization*. Heidelberg: Springer, 2003, 270–278.

23 Mirkin B, Arabie P, Hubert LJ. Additive two-mode clustering: the error-variance approach revisited. *Journal of Classification* 1995; **12**: 243–63.

24 Carroll JD, Arabie P. Multidimensional scaling. *Annual Review of Psychology* 1980; **31**: 607–49.

25 Tucker LR. The extension of factor analysis to three-dimensional matrices. In Frederiksen N, Gulliksen H, eds. *Contributions to mathematical psychology*. New York: Holt, Rinehart and Winston, 1964, 109–127.

26 Frege G. *Grundgesetze der Arithmetik, begrifflich abgeleitet*, Band II. Jena: Verlag Hermann Pohle, 1903.

27 Braverman EM. Methods for extremal grouping of the variables and the problem of finding important factors. *Automation and Remote Control* 1970; **31**: 123–32.

28 Bock HH. Probability models and hypotheses testing in partitioning cluster analysis. In Arabie P, Hubert LJ, De Soete G, eds. *Clustering and classification.* River Edge, NJ: World Scientific, 1996.

29 Lambert JM, Williams WT. Multivariate methods in plant ecology. IV. Nodal analysis. *Journal of Ecology* 1962; **50**: 775–802.

30 Williams WT, Lambert JM. Nodal analysis of associated populations. *Nature* 1961; **191**: 202.

31 Breiger RL, Boorman SA, Arabie P. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology* 1975; **12**: 328–83.

32 Arabie P, Boorman SA, Levitt PR. Constructing blockmodels: How and why. *Journal of Mathematical Psychology* 1978; **17**: 21–63.

33 Noma E, Smith DR. Benchmark for the blocking of sociometric data. *Psychological Bulletin* 1985; **97**: 583–91.

34 Arabie P, Schleutermann S, Daws J, Hubert LJ. Marketing applications of sequencing and partitioning of nonsymmetric and/or two-mode matrices. In Gaul W, Schader M, eds. *Data, expert knowledge and decisions*. Berlin: Springer-Verlag, 1988, 215–224.

35 Arabie P, Hubert LJ, Schleutermann S. Blockmodels from the bond energy approach. *Social Networks* 1990; **12**: 99–126.

36 Arabie P, Hubert LJ. The bond energy algorithm revisited. *IEEE Transactions on Systems, Man, and Cybernetics* 1991; **20**: 268–74.

37 Marcotorchino F. Block seriation problems: a unified approach. *Journal of Applied Stochastical Models and Data Analysis* 1987; **3**: 73–93.

38 Govaert G. Algorithme de classification d'un tableau de contingence. *Premières journées internationales analyse des données et informatique (Versailles 1977)*. Paris: CNRS, 1980, 487–500.

39 Govaert G. Classification simultanée de tableaux binaires. In Diday E, Jambu M, Lebart L, Pages J, Tomassone R, eds. *Data analysis and informatics 3*. Amsterdam: North Holland, 1984, 223–236.

40 Govaert G. Simultaneous clustering of rows and columns. *Control and Cybernetics* 1995; **24**: 437–58.

41 DeSarbo WS. GENNCLUS: new models for general nonhierarchical clustering analysis. *Psychometrika* 1982; **47**: 449–75.

42 Gaul W, Schader M. A new algorithm for two-mode clustering. In Bock H-H, Polasek W, eds. *Data analysis and information systems*. Heidelberg: Springer, 1996, 15–23.

43 Baier D, Gaul W, Schader M. Two-mode overlapping clustering with applications in simultaneous benefit segmentation and market structuring. In Klar R, Opitz O, eds. *Classification and knowledge organization.* Heidelberg: Springer, 1997, 557–566.

44 Vichi M. Double k-means clustering for simultaneous classification of objects and variables. In Borra S, Rocchi R, Schader M, eds. *Advances in classification and data analysis. Studies in classification, data analysis, and knowledge organization.* Heidelberg: Springer, 2001, 43–52.

45 Trejos J, Castillo W. Simulated annealing optimization for two-mode partitioning. In Gaul W, Decker R, eds. *Classification and information at the turn of the millenium*. Heidelberg: Springer, 2000, 135–142.

46 Castillo W, Trejos J. Two-mode partitioning: review of methods and application of tabu search. In Jajuga K, Sokolowski A, Bock H-H, eds. *Classification, clustering, and related topics. Recent advances and applications. Studies in classification, data analysis, and knowledge organization.* Heidelberg: Springer-Verlag, 2002, 43–51.

47 Hansohm J. Two-mode clustering with genetic algorithms. In Gaul W, Ritter G, eds. *Classification, automation, and new media. Studies in classification, data analysis, and knowledge organization.* Heidelberg: Springer, 2002, 87–93.

48 Bock H-H. Simultaneous clustering of objects and variables. In Tomassone R, ed. *Analyse des données et informatique.* Le Chesnay, France: INRIA, 1979, 187–204.

49 Greenacre MJ. Clustering the rows and columns of a contingency table. *Journal of Classification* 1988; **5**: 39–51.

50 Bock H-H. Convexity-based clustering criteria: a new approach. *Recor's Lecture presented at the Academy of Economy of Cracow*, Poland, 2000.

51 Bock H-H. Two-way clustering for contingency tables: Maximizing a dependence measure. In Schader M, Gaul W, Vichi M, eds. *Between data science and applied data analysis.* Heidelberg-Berlin: Springer Verlag, 2003, 143–54.

52 Bock H-H. Convexity-based clustering criteria: theory, algorithms, and applications in statistics. *Statistical Methods and Applications* 2003; **12**: 293–317.

53 Bock H-H. A clustering algorithm for choosing optimal classes for the chi-squared test. *Bulletin of the International Statistical Institute, 44th session.* Madrid, 1983.

54 Pötzelberger K, Strasser H. Clustering and quantization by MSP-partitions. *Statistics and Decisions* 2001; **19**: 331–71.

55 Ciok A. Discretization as a tool in cluster analysis. In Rizzi A, Vichi M, Bock H-H, eds. *Advances in data science and classification.* Heidelberg: Springer, 1998, 349–54.

56 DeSarbo WS, Fong DKH, Liechty J. A hierarchical Bayesian procedure for two-mode cluster analysis. *Paper presented at the 27th Annual Meeting of the Gesellschaft für Klassifikation.* Brandenburg: University of Technology Cottbus, 2003.

57 Govaert G, Nadif M. Clustering with block mixture models. *Pattern Recognition* 2003; **36**: 463–73.

58 Hartigan JA. Bloc voting in the United States Senate. *Journal of Classification* 2000; **17**: 29–49.

59 Hartigan JA. Partition models. *Communications in Statistics* 1990; **19**: 2745–56.

60 Duffy DE, Quiroz AJ. A permutation-based algorithm for block clustering. *Journal of Classification* 1991; **8**: 65–91.

61 Furnas GW. *Objects and their features: the metric analysis of two-class data.* Doctoral dissertation. Stanford University, Stanford, CA, 1980.

62 De Soete G, Carroll JD. Tree and other network models for representing proximity data. In Arabie P, Hubert LJ, De Soete G, eds. *Clustering and Classification.* River Edge, NJ: World Scientific, 1996, 157–197.

63 Tversky A. Features of similarity. *Psychological Review* 1977; **84**: 327–52.

64 Eckes T, Orlik P. An agglomerative method for two-mode hierarchical clustering. In Bock H-H, Ihm P, eds. *Classification, data analysis, and knowledge organization.* Berlin: Springer-Verlag, 1991, 3–8.

65 Eckes T. A two-mode clustering study of situations and their features. In Opitz O, Lausen B, Klar R, eds. *Information and classification.* New York: Springer-Verlag, 1993, 510–517.

66 Castillo W, Trejos J. Recurrence properties in two-mode hierarchical clustering. In Decker R, Gaul W, eds. *Classification and information processing at the turn of the millennium.* Heidelberg: Springer, 2000, 68–73.

67 Lance GN, Williams WT. A general theory of classification sorting strategies. *Computer Journal* 1967; **9**: 373–80.

68 Schwaiger M. Two-mode classification in advertising research. In Klar R, Opitz, O, eds. *Classification and knowledge organization.* Heidelberg: Springer, 1997, 597–603.

69 Espejo E, Gaul W. Two-mode hierarchical clustering as an instrument for marketing research. In Gaul W, Schader M, eds. *Classification as a tool of research.* Amsterdam: Elsevier/North-Holland, 1986, 121–128.

70 De Soete G, DeSarbo WS, Furnas GW, Carroll JD. The estimation of ultrametric and path length trees from rectangular proximity data. *Psychometrika* 1984; **49**: 289–310.

71 Hubert LJ, Arabie P. Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British Journal of Mathematical and Statistical Psychology* 1995; **48**: 281–317.

72 Hartigan JA. Modal blocks in dentition of west coast mammals. *Systematic Zoology* 1976; **25**: 149–60.

73 Shepard RN, Arabie P. Additive clustering representation of similarities as combinations of discrete overlapping properties. *Psychological Review* 1979; **86**: 87–123.

74 Carroll JD, Chaturvedi A. A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data. In Luce RD, D'Zmura M, Hoffman DD, Iverson GJ, Romney AK, eds. *Geometric representations of perceptual phenomena.* Mahwah: Erlbaum, 1995, 295–318.

75 Mickey MR, Mundle P, Engelman L. Boolean factor analysis. In Dixon WJ, ed. *BMDP statistical software manual.* Berkeley, CA: University of California Press, 1983, 538–45.

76 De Boeck P, Rosenberg S. Hierarchical classes: model and data analysis. *Psychometrika* 1988; **53**: 361–81.

77 Van Mechelen I, Rosenberg, S, De Boeck, P. On hierarchies and hierarchical classes models. In Mirkin B, McMorris FR, Roberts FS, Rhetsky A, eds. *Mathematical hierarchies and biology.* Providence: American Mathematical Society, 1997, 291–98.

78 Van Mechelen I, De Boeck P, Rosenberg, S. The conjunctive model of hierarchical classes. *Psychometrika* 1995; **60**: 505–21.

79 Leenen I, Van Mechelen I. An evaluation of two algorithms for hierarchical classes analysis. *Journal of Classification* 2001; **18**: 57–80.

80 Barbut M, Monjardet B. *Ordre et classification.* Paris: Classiques Hachette, 1970.

81 Ganter B, Wille R. Formal concept analysis: mathematical foundations. Berlin: Springer, 1999.

82 Stahringer S, Wille R. Conceptual clustering via convex-ordinal structures. In Opitz O, Lausen B, Klar R, eds. *Information and Classification.* New York: Springer-Verlag, 1993, 85–98.

83 Krolak-Schwerdt S, Orlik P. Direct clustering of a two-mode binary data matrix. *Arbeiten der Fachrichtung Psychologie.* Saarbrücken: Universität des Saarlandes, 1998.

84 Van Mechelen I. Approximate Galois lattices of formal concepts. In Opitz O, Lausen B, Klar R, eds. *Information and classification.* New York: Springer-Verlag, 1993, 108–112.

85 Mezzich JE, Solomon H. *Taxonomy and behavioral science: comparative performance of grouping methods.* London: Academic Press, 1980.

86 Gara M, Rosenberg S, Goldberg L. DSM-IIIR as a taxonomy: a cluster analysis of diagnoses and symptoms. *Journal of Nervous and Mental Disease* 1992; **180**: 11–19.

87 Ceulemans E. An algorithm for the HICLAS-R model. In Schader M, Gaul W, Vichi M, eds. *Between data science and applied data analysis.* Heidelberg, Berlin: Springer Verlag, 2003, 173–181.