Probabilistic Two-way Clustering Approaches with Emphasis on the Maximum Interaction Criterion

Hans-Hermann Bock

Abstract We consider the problem of simultaneously and optimally clustering the rows and columns of a real-valued $I \times J$ data matrix $X = (x_{ij})$ by corresponding row and columns partitions $\mathscr{A} = (A_1, ..., A_m)$ and $\mathscr{B} = (B_1, ..., B_n)$, with given m and n. We emphasize the need to base the clustering method on a probabilistic model for the data and then to use standard methods from statistics (e.g., maximum likelihood, divergence) to characterize optimum two-way classifications. We survey some clustering criteria and algorithms proposed in the literature for various data types. Special emphasis is given to the maximum interaction clustering criterion proposed by the author in 1980. It can be shown that it results as the maximum likelihood clustering method under a two-way ANOVA model (with individual main effects, but cluster-specific interactions). After a simple data transformation (double-centering) well-known two-way SSQ clustering algorithms can directly be used for maximization.

Hans-Hermann Bock
Institute of Statistics, RWTH Aachen University, Vorder-Winterbach 36, 77794 Lautenbach, Germany,
✉ bock@stochastik.rwth-aachen.de

ARCHIVES OF DATA SCIENCE, SERIES A
KIT SCIENTIFIC PUBLISHING
Vol. 1, No. 1, S. 3–20, 2016

DOI 10.5445/KSP/1000058747/01
ISSN 2363-9881

1 Two-way clustering problems

Two-way clustering means clustering, simultaneously, the rows and columns of a data matrix $X = (x_{ij})_{I\times J}$. Synonymns are bi-clustering, co-clustering, or block clustering. In practice, two-way clustering problems occur, e.g.,

- in microbiology (microarray measurements for I genes and J different times, situations, or tissues); see, e.g., Martella et al (2008), Cheng and Church (2000), Madeira and Oliveira (2004), Martella et al (2011), Martella and Vichi (2012), Turner et al (2005)
- in marketing (purchase data for I consumers described by J social characteristics); see, e.g., Baier et al (1997), Arabie et al (1988)
- in documentation (I documents or e-mails described by presence/absence of J keywords); see, e.g., Dhillon et al (2003), Banerjee et al (2007), Li and Zha (2006), Cho et al (2004), Cho and Dhillon (2008).

Many two-way clustering methods have been proposed since the beginning of clustering activities in the 1970s (recent surveys were given by Van Mechelen et al, 2004; Madeira and Oliveira, 2004; Charrad and Ben Ahmed, 2011; Vichi, 2012; Govaert and Nadif, 2013), but the possibility to record automatically huge sets of data in various application fields has meanwhile increased the importance of two-way clustering for an adequate and informative analysis of data.

In this paper we consider a real-valued data matrix $X = (x_{ij})_{I\times J}$ with I rows, J columns and try to find an m-partition $\mathscr{A} = (A_1,...,A_m)$ of the row set $\mathscr{I} = \{1,...,I\}$ with m classes, and an n-partition $\mathscr{B} = (B_1,...,B_n)$ of the column set $\mathscr{C} = \{1,...,J\}$ with n classes, such that the joint $m\cdot n$-partition $\mathscr{A}\times\mathscr{B} = \{A_r \times B_s | r = 1,...,m, s = 1,...,n\}$ of the set of pairs $\{(i,j)|i \in \mathscr{I}, j \in \mathscr{J}\}$ (cells of the matrix X) together with a suitable parametric characterization of the classes fits, approximates or reproduces optimally the hidden row by column structure (if any) in the given data matrix X. Obviously, such a formulation requires the specification of some “structure" that should be reconstructed from the data, and some optimality criterion that should be optimized. The multitude of proposed two-way clustering algorithms can be largely explained by the great number of choices for “structure" and “optimality".

We emphasize here the probabilistic approach where “structure" is described by a parametric and block-specific probability distribution for the data X_{ij}. Then, generally, the parameter estimates as well as the bi-clustering $(\mathscr{A},\mathscr{B})$ are obtained by the maximum-likelihood (m.l.) approach. Thereby, the choice of a distributional model is highly dependent on the way in which the data were obtained and on their interpretation as measurement values, associations, frequencies, indicators, etc. In this respect we will consider

- association-type data for a two-mode data matrix (Sect. 2
- measurement-type values x_{ij} with categorical factor levels i, j (Sect. 3)
- frequency-type values N_{ij} with factor levels i, j (contingency table; Sect. 4)
- object by variable measurements x_{ij} (classical data matrix; Sect. 5)

and provide some exemplary probabilistic clustering approaches. For binary variables we refer, e.g., to Govaert and Nadif (2005); Li (2005); Govaert and Nadif (2007, 2008, 2013) and Nadif and Govaert (2010).

Note that we will not comment here on the choice of the numbers m,n of classes (see, e.g., Schepers et al, 2008) and will present only the so-called "fixed-partition" or "classification likelihood" approaches (see, e.g., Bock, 1996a,b). Alternatively, probabilistic clustering approaches can also be formulated in terms of mixture models ('random-partition" approach) resulting in EM-type algorithms and fuzzy bi-partitions in the form of posterior distributions (see, e.g., Govaert, 1995; Govaert and Nadif, 2005, 2003, 2008, 2010; Bocci et al, 2006; Li and Zha, 2006; Martella et al, 2008, 2011). Other approaches use row- and column-wise hierarchical clusterings or try to cover the set of IJ matrix cells with suitably weighted, possibly overlapping "homogenous blocks" $A \times B$ such as *plaid methods* (described by Lazzeroni and Owen, 2002; Turner et al, 2005) or *additive clustering* (as in Shepard and Arabie, 1979; Mirkin et al, 1995; Wilderjans et al, 2013). See also the articles on multi-mode clustering in the Special Issue on "Statistical learning methods including dimension reduction" of the journal "Computational Statistics and Data Analysis" (vol. 52, 2007, edited by H.-H. Bock and M. Vichi).

2 Clustering for association-type data

In this section we suppose that the data x_{ij} represent association values that measure how "close", "associated", or "interrelated" row i is to column j. Also we assume a two-mode case, i.e., rows and columns refer to different sets (such as customers and products, genes and time points, respectively). In this case a classical two-way clustering criterion is provided by the SSQ:

$$g(\mathscr{A},\mathscr{B},\mu) := \sum_{r=1}^{m}\sum_{s=1}^{n}\sum_{i\in A_r}\sum_{j\in B_s} ||x_{ij}-\mu_{rs}||^2 \to min_{\mathscr{A},\mathscr{B},\mu} \qquad (1)$$

where $\mu_{rs} \in R$ is a block-specific prototype value and μ the set of these values[1] (Bock, 1980). This criterion amounts to approximating the given data matrix X by an “ideal" block-matrix $\widetilde{X}_{I\times J}$ with the same value μ_{rs} in all cells of a block (bicluster) $A_r \times B_s$ (for all r,s). Given that partial minimization with respect to μ leads to the average values $\hat{\mu}_{rs} = \bar{x}_{A_r\times B_s}$ in the blocks $A_r \times B_s$ of X, the criterion (1) is equivalent to the following *SSQ clustering criterion*:

$$Q_{min}(\mathscr{A},\mathscr{B};X) := \sum_{r=1}^{m}\sum_{s=1}^{n}\sum_{i\in A_r}\sum_{j\in B_s} ||x_{ij}-\bar{x}_{A_r\times B_s}||^2 \to \min_{\mathscr{A},\mathscr{B}} \qquad (2)$$

and to

$$k(\mathscr{A},\mathscr{B};X) := \sum_{r=1}^{m}\sum_{s=1}^{n} |A_r|\cdot|B_s|\cdot||\bar{x}_{A_r\times B_s}||^2 \to \max_{\mathscr{A},\mathscr{B}}. \qquad (3)$$

In order to optimize these clustering criteria many algorithms (e.g., double k-means) have been proposed; see, e.g., Bock (1980); Gaul and Schader (1996); Baier et al (1997); Hansohm (2002); Vichi (2001); Castillo and Trejos (2002); Cho et al (2004); Cho and Dhillon (2008); Rocci and Vichi (2008); Van Rosmalen et al (2009); Schepers and Hofmans (2009); Martella and Vichi (2012)

3 Clustering for factorial designs

In this section we consider the case where all data values x_{ij} are measurements of the same target variable which, however, depends on two categorical factors U (rows) and V (columns) with categories in $\mathscr{I} = \{1,...,I\}$ and $\mathscr{J} = \{1,...,J\}$, respectively. For example, in a diet experiment with many persons, U might be the initial BMI (discretized body mass index, $I = 30$, say) of a person, V the type of diet that this person applies (with $J = 15$ types, say), and x_{ij} the average loss of weight after a four-weeks diet for all persons with $U = i$ and $V = j$. Assuming a complete factorial design (i.e., observations were made for

[1] $||x||$ means the absolute value $|x|$ for $x \in R^1$ and the Euclidean norm for multivariate data (see Remark 2). For a set A, $|A|$ means the number of elements of A.

all IJ combinations $(i,j) \in \mathscr{I} \times \mathscr{J}$) the clustering problem consists in finding (a given number $m = 6$, say, of) BMI classes $A_1,...,A_m$ and (a given number $n = 4$, say, of) diet classes $B_1,...,B_n$ that best describe the data. In this way, the large number of categories can be reduced to a smaller and handy number of category classes or "types".

Classical statistics analyzes such two-way configurations by ANOVA models with random variables X_{ij} that are additively obtained from a total mean, row and column main effects, interaction terms, and normal errors. In the clustering framework we consider two such models: one with individual main effects, and one with class-specific main effects. It appears that only the first one provides new insights while the second one falls back to the criterion (2).

3.1 ANOVA clustering model with individual main effects

Here we assume that the existence of a hidden bi-clustering is exclusively caused by block-specific interaction terms while main effects do not contribute to the clustering aspect. In the framework of ANOVA this amounts to suppose that X_{ij} are given, for a fixed bi-partition $(\mathscr{A},\mathscr{B})$, by the additive composition:

$$X_{ij} = c + a_i + b_j + \gamma_{rs} + e_{ij} \qquad i \in A_r, j \in B_s, r = 1,...,m, s = 1,...,n. \quad (4)$$

Here c is a fixed mean value, a_i the *individual* main effect of category i of U, b_j the *individual* main effect of category j of V, and γ_{rs} the *class-specific* interaction effect; the latter one is the same for all pairs (i,j) in the bicluster $A_r \times B_s$. The e_{ij} are independent random error terms with $e_{ij} \sim \mathscr{N}(0,\sigma^2)$ where we consider σ^2 to be known here (but see Remark 1). In order to attain identifiability of parameters, the following zero-means normalization is introduced:

$$\begin{array}{lll} \bar{a}_\bullet := \sum_{i=1}^I a_i/I = 0, & \bar{b}_\bullet := \sum_{j=1}^J b_j/J = 0, & \\ \bar{\gamma}_{\bullet,s} := \sum_{r=1}^m |A_r| \cdot \gamma_{rs}/I = 0, & \bar{\gamma}_{r,\bullet} := \sum_{s=1}^n |B_s| \cdot \gamma_{rs}/J = 0 & \text{for all } r,s. \end{array}$$

For estimating the unknown parameters c, a_i, b_j, γ_{rs} and the unknown $(\mathscr{A},\mathscr{B})$ we use the m.l. approach. Due to the normality assumptions this amounts to minimizing the SSQ:

$$\widetilde{Q}(c,a,b,\gamma,\mathscr{A},\mathscr{B}) := \sum_{r=1}^m \sum_{s=1}^n \sum_{i \in A_r} \sum_{j \in B_s} ||x_{ij} - c - a_i - b_j - \gamma_{rs}||^2 \to \min_{c,a,b,\gamma,\mathscr{A},\mathscr{B}} \quad (5)$$

After some algebraic manipulations (or using derivatives) we obtain, for a fixed bi-partition $(\mathscr{A},\mathscr{B})$, the following m.l. estimates:

$\hat{c} = \bar{x}_{\bullet,\bullet}$	overall mean
$\hat{a}_i = \bar{x}_{i,\bullet} - \bar{x}_{\bullet,\bullet}$ and $\hat{b}_j = \bar{x}_{\bullet,j} - \bar{x}_{\bullet,\bullet}$	individual main effects
$\hat{\gamma}_{rs} = \bar{x}_{A_r\times B_s} - \bar{x}_{A_r,\bullet} - \bar{x}_{\bullet,B_s} + \bar{x}_{\bullet,\bullet}$	class-specific interaction effects.

Inserting these estimates into (5) yields the clustering criterion:

$$\widetilde{Q}_{min}(\mathscr{A},\mathscr{B}) := \sum_{r=1}^{m}\sum_{s=1}^{n}\sum_{(i,j)\in A_r\times B_s}(x_{ij} - \hat{\mu} - \hat{a}_i - \hat{b}_j - \hat{\gamma}_{rs})^2 \quad \to \quad \min_{\mathscr{A},\mathscr{B}} \tag{6}$$

that can be shown, by algebraic transformations (see Bock, 1980; Schepers et al, 2013), to be equivalent to the following *maximum interaction clustering criterion*:

$$\begin{aligned} G(\mathscr{A},\mathscr{B};X) &:= \sum_{r=1}^{m}\sum_{s=1}^{n} |A_r|\cdot|B_s|\cdot|\hat{\gamma}_{rs}^{(X)}|^2 \\ &= \sum_{r=1}^{m}\sum_{s=1}^{n} |A_r|\cdot|B_s|\cdot(\bar{x}_{A_r\times B_s} - \bar{x}_{A_r,\bullet} - \bar{x}_{\bullet,B_s} + \bar{x}_{\bullet,\bullet})^2 \to \max_{\mathscr{A},\mathscr{B}} \end{aligned} \tag{7}$$

where we have flagged $\hat{\gamma}_{rs}^{(X)}$ by the superscript X in order to emphasize the corresponding data matrix X.

This clustering criterion was proposed by Bock (1980) on empirical grounds. The previous argumentation shows that it derives from the probabilistic factorial ANOVA approach (4). In Sect. 4 we will show that its minimization can be easily performed by the algorithms that were developed for the SSQ cluster criterion (2); so no specific algorithms have to be developed for (7).

Remark 1: It can easily be shown that the criterion (7) results as the m.l. clustering criterion also in the case of an unknown variance σ^2.

Remark 2: In case of vector-valued variables X_{ij} and observations $x_{ij} \in R^p$ the ANOVA model (4) must be formulated with p-dimensional effects c, a_i, b_j, γ_{rs} and $e_{ij} \sim \mathscr{N}_p(0, I_p)$. For this p-dimensional version the m.l. clustering approach yields the same clustering criteria as before (in particular, the maximum interaction criterion (7)) where $||...||$ now is the Euclidean norm in R^p.

3.2 ANOVA clustering model with class-specific main effects

We may wonder what happens if we assume that in the ANOVA model (4) not only the interactions, but also the main effects are class-specific. This amounts to the additive model

$$X_{ij} = \mu_{rs} + e_{ij} = c + \alpha_r + \beta_s + \gamma_{rs} + e_{ij} \quad i \in A_r, j \in B_s, r = 1,...,m, s = 1,...,n \tag{8}$$

with class-specific “block prototypes" $\mu_{rs} = c + \alpha_r + \beta_s + \gamma_{rs}$, typically with a zero-mean standardization for the effects $\alpha_r, \beta_s, \gamma_{rs}$. Note that for given $\{\mu_{rs}\}$ the standardized effects are uniquely determined by $c = \bar{\mu}_{\bullet,\bullet}$, $\alpha_r := \bar{\mu}_{A_r,\bullet} - \bar{\mu}_{\bullet,\bullet}$, $\beta_s = \bar{\mu}_{\bullet,B_s} - \bar{\mu}_{\bullet,\bullet}$ and $\gamma_{rs} = \bar{\mu}_{A_r,B_s} - \bar{\mu}_{A_r,\bullet} - \bar{\mu}_{\bullet,B_s} + \bar{\mu}_{\bullet,\bullet}$ such that the parameter sets $\{\mu_{rs}\}$ and $\{c, a_r, b_s, \gamma_{rs}\}$ are uniquely determined by each other. Therefore only the μ_{rs} must be estimated.

Due to the normality assumption m.l. clustering is here equivalent to minimizing the total SSQ (1) with respect to $\{\mu_{rs}\}$ and $(\mathscr{A}, \mathscr{B})$. Therefore all statements of Sect. 2 apply and insofar also the clustering criteria (2) and (3) are justified by a probabilistic model (Bock, 1980).

3.3 Maximizing the interaction criterion

Surprisingly it appears that the interaction criterion $G(\mathscr{A}, \mathscr{B}; X)$, (7), can be (approximately) maximized by the same algorithms that have been developed for minimizing the SSQ criterion $Q_{min}(\mathscr{A}, \mathscr{B}; Y)$, (2), if the original data matrix X is suitably transformed before (see also Bock, 1980). In fact:

Theorem 1. *Maximizing the interaction criterion $G(\mathscr{A}, \mathscr{B}; X)$ from (7) is equivalent to minimizing the SSQ clustering criterion $Q_{min}(\mathscr{A}, \mathscr{B}; Y)$ from (2) where the data matrix X has been replaced by the double-centered matrix $Y = (y_{ij})_{I \times J}$ with entries*

$$y_{ij} := x_{ij} - \bar{x}_{i,\bullet} - \bar{x}_{\bullet,j} + \bar{x}_{\bullet,\bullet} \qquad \textit{for all } i, j.$$

Proof. It is easily seen that for all r,s:

$$\bar{y}_{A_r\times B_s} \;=\; \bar{x}_{A_r\times B_s} - \bar{x}_{A_r,\bullet} - \bar{x}_{\bullet,B_s} + \bar{x}_{\bullet,\bullet} \;=\; \hat{\gamma}_{rs}^{(X)}.$$

Therefore the interaction criterion $G(\mathscr{A},\mathscr{B};X)$ is identical to the criterion $k(\mathscr{A},\mathscr{B};Y)$ from (3). On the other hand, the well-known decomposition formula

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}||y_{ij}||^2 &= \underbrace{\sum_{r=1}^{m}\sum_{s=1}^{n}\sum_{\substack{(i,j)\in\\ A_r\times B_s}}||y_{ij}-\bar{y}_{A_r\times B_s}||^2}_{} + \underbrace{\sum_{r=1}^{m}\sum_{s=1}^{n}|A_r|\cdot|B_s|\cdot||\bar{y}_{A_r\times B_s}||^2}_{} \\ &= \qquad Q_{min}(\mathscr{A},\mathscr{B};Y) \qquad + \qquad k(\mathscr{A},\mathscr{B};Y). \end{aligned} \qquad (9)$$

(where the left hand side is constant with respect to $\mathscr{A},\mathscr{B}$) shows that maximizing the criterion $k(\mathscr{A},\mathscr{B};Y)$ is equivalent to minimizing the SSQ criterion $Q_{min}(\mathscr{A},\mathscr{B};Y)$ for the double-centered matrix Y. qed

4 Two-way clustering for a contingency table

In this section we consider again a two-way factorial design with two categorical characteristics U and V as in Sect. 3, but here we assume that the entries x_{ij} of the data matrix X are counts N_{ij} and write $X = \mathscr{N} = (N_{ij})_{I\times J}$ in this case. As an example we may consider the N clients (contracts) of a car insurance company, characterized by the profession U of the client and the brand V of the insured car. Then N_{ij} is the number of clients with profession i and car make j. For the company it can make sense to reduce the large numbers of categories I and J to a smaller number m of (profession) classes A_r and a smaller number n of (brand) classes B_s such that profession classes are, on the average, most predictive for the brand class of a client, i.e., with a maximum interaction between both. The resulting classes A_r, B_s and biclusters $A_r \times B_s$ might be the basis for calculating adequate insurance premiums.

In contrast to Sect. 3 where normal distributions were involved, the new scenario is modeled by a random sample of N items (clients) such that N_{ij} is the number of items assigned to the category combination (i,j) (with $\sum_{ij} N_{ij} = N$). Then $\mathscr{N} = (N_{ij})$ has a polynomial distribution $\mathscr{P}ol(N;(p_{ij})_{I\times J})$ with unknown cell probabilities p_{ij} which are typically estimated by $\hat{p}_{ij} := N_{ij}/N$.

In this framework “independence among row and column classes" is modeled by the “hypothesis" H_0:

$$P(A_r \times B_s) = P_U(A_r) \cdot P_V(B_s) \qquad \text{for all } r, s$$

with $P(A_r \times B_s) := \sum_{i \in A_r} \sum_{j \in B_s} p_{ij}$, $P_U(A_r) := \sum_{i \in A_r} \sum_{j=1}^{J} p_{ij}$,
$P_V(B_s) := \sum_{i=1}^{I} \sum_{j \in B_s} p_{ij}$, and can be tested, for a fixed bi-partition $(\mathscr{A}, \mathscr{B})$, by the classical χ^2 test. On the other hand, the contrasting idea of “maximum interaction between row and column classes" is interpreted here in the way that the χ^2 test is maximally significant for rejecting H_0, i.e., that the χ^2 test statistics, termed χ^2 *clustering criterion*

$$C(\mathscr{A}, \mathscr{B}) := \sum_{r=1}^{m} \sum_{s=1}^{n} \frac{(\hat{P}(A_r \times B_s) - \hat{P}_U(A_r) \cdot \hat{P}_V(B_s))^2}{\hat{P}_U(A_r) \cdot \hat{P}_V(B_s)} \to \max_{\mathscr{A}, \mathscr{B}} \tag{10}$$

is maximal with respect to the bi-partition $(\mathscr{A}, \mathscr{B})$. Here $\hat{P}$ means the m.l. estimate for the probability distribution P, e.g. with $\hat{P}_{U,V}(A_r \times B_s) = \sum_{i \in A_r} \sum_{j \in B_s} \hat{p}_{ij}$ $= \sum_{i \in A_r} \sum_{j \in B_s} N_{ij}/N$.

In a more general context we note that the χ^2 criterion (10) results as a special case (for $\phi(\lambda) := (\lambda - 1)^2$) from the classical ϕ-divergence measure by Csiszár:

$$C_\phi(\mathscr{A}, \mathscr{B}) := \sum_{r=1}^{m} \sum_{s=1}^{n} \hat{P}_U(A_r) \hat{P}_V(B_s) \cdot \phi\left(\frac{\hat{P}(A_r \times B_s)}{\hat{P}_U(A_r) \hat{P}_V(B_s)}\right) \to \max_{\mathscr{A}, \mathscr{B}} \tag{11}$$

where ϕ is an arbitrary convex function. This *divergence clustering criterion* measures the deviation between the observed probability distribution $\hat{P}$ and the product distribution $\hat{P}_U \cdot \hat{P}_V$ for a given biclustering $(\mathscr{A}, \mathscr{B})$. For $\phi(\lambda) = -\log \lambda$ a Kullback-Leibler clustering criterion results. These criteria have been proposed for clustering by Bock (1983, 1992, 2003, 2004), Celeux et al (1989, χ^2 criterion), Dhillon et al (2003) and Banerjee et al (2005, 2007). Note that the usage of the χ^2 criterion can be justified by theoretical considerations in terms of maximum power, Bahadur efficiency etc. of the χ^2 test (Bock, 1992).

In order to minimize the divergence criterion we may use the classical alternating maximization scheme (*generalized double k-means*): Choose an initial bipartition $\mathscr{A}^{(0)}, \mathscr{B}^{(0)}$ and then alternate between (i) partial maximization with respect to the row partition $\mathscr{A}$ (for fixed $\mathscr{B}$) and (ii) partial maximization with respect to the column partition $\mathscr{B}$ (for fixed $\mathscr{A}$). In order to conduct these partial minimization steps Bock (1992, 2003, 2004) has proposed a k-means-type algorithm that uses class-specific tangents (subgradients) of the convex function ϕ (instead of class means as in the classical SSQ case) and was therefore termed *k-tangent algorithm*. See also Dhillon et al (2003) and Banerjee et al (2005, 2007). For a mixture-type approach see Govaert and Nadif (2010, 2013).

5 Two-way clustering for an object by variable matrix

In the previous sections clustering of rows and columns of the data matrix $X = (x_{ij})_{I\times J}$ was performed in a symmetrical way such that the roles of rows and columns could have been reversed without changing the results. This is different in the case of an object by variable data matrix since, e.g., objects will be independently sampled while variables might be more or less dependent. Also the motivations for grouping objects and variables are different: objects are assembled in groups because they are supposed to behave similarly (with respect to all variables) whereas variables from the same group are supposed to be dependent from each other while independence may hold for variables of different groups. In this last section we sketch two approaches for modeling bi-partition structures for X in the case of I objects and J continuous variables. For more information see, e.g., Vichi (2012); Nadif and Govaert (2010); Govaert and Nadif (2013).

In a probabilistic framework the rows $x_i = (x_{i1},...,x_{iJ})'$ of X are considered as a sample of I independent random (column) vectors $X_i = (X_{i1},...,X_{iJ})'$ with a distribution that depends on the group A_r of $\mathscr{A} = (A_1,...,A_m)$ to which object i belongs to. Any clustering $\mathscr{B} = (B_1,...,B_n)$ of the set of columns $\mathscr{J}$ (with group sizes $b_s := |B_s|$, $s = 1,...,n$, $\sum_s b_s = J$) is supposed to split the set $\mathscr{J}$ of variables into n mutually independent groups of variables. This also amounts to splitting X_i into n subvectors $X_{i,B_1},...,X_{i,B_n}$ such that $X_{i,B_s} \in R^{b_s}$ comprizes the components X_{ij} of X_i that belong to class B_s. For notational convenience we assume here that the ordering of components in X_i is such that all classes $B_1,...,B_n$ comprize contiguous sets of variables $j \in \mathscr{J}$ such that $X_i = (X'_{i,B_1},...,X'_{i,B_n})'$.

A first clustering model is based on the J-dimensional normal distribution:

$$X_i := \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iJ} \end{pmatrix} = \begin{pmatrix} X_{i,B_1} \\ \vdots \\ X_{i,B_n} \end{pmatrix} \sim \mathscr{N}_J(\mu^{(r)}(\mathscr{B}); \Sigma^{(r)}(\mathscr{B})) \quad \text{for } i \in A_r \qquad (12)$$

($r = 1,...,m$) where object classes A_r are characterized by class-specific and partitioned expectations $\mu^{(r)}(\mathscr{B}) \in R^J$ and $J \times J$ covariance matrices $\Sigma^{(r)}(\mathscr{B})$ according to

$$\mu^{(r)}(\mathscr{B}) = \begin{pmatrix} \mu_{r,B_1} \\ \vdots \\ \mu_{r,B_n} \end{pmatrix} \qquad \Sigma^{(r)}(\mathscr{B}) = diag(\Sigma_{11}^{(r)}, \cdots \Sigma_{nn}^{(r)}) \qquad (13)$$

In particular, we then have, for all $i \in A_r$, that $X_{i,B_s} \sim \mathscr{N}_{b_s}(\mu_{r,B_s}, \Sigma_{ss}^{(r)})$ with independent subvectors X_{i,B_s}, X_{i,B_t} for different column classes B_s and B_t.

While, in principle, m.l. clustering might be possible for this general case, practical applications may concentrate on more parsimonious covariance models, e.g.:

- with independent variables within each group: $\Sigma_{ss}^{(r)} = {\sigma_s^{(r)}}^2 I_{b_s}$ for all s (and then, a fortiori, independence among all J variables);
- with the same variances in all object classes A_r: ${\sigma_s^{(r)}}^2 = \sigma_s^2$ for all r and s;
- with the same variances $\sigma_1^2 = \cdots = \sigma_n^2$ for all groups B_s (then variable groups differ only by the expectation vectors μ_{r,B_s}).

A related mixture model approach is described, e.g., by Nadif and Govaert (2010).

A second modeling approach is based on characteristic subspaces for the variables in B_s, but is only briefly sketched here in a simple case. Let us denote the J column variables of X by $Y_1,...,Y_J$. We start from the assumption that within each column class B_s, the corresponding random vector Y_{B_s} (that corresponds to the subvector X_{i,B_s} in the matrix X) is generated by a T-dimensional random vector $U^{(s)} := (U_1^{(s)},...,U_T^{(s)})'$ such that $Y_{B_s} = \alpha^{(s)} + \sum_{t=1}^T \beta_t^{(s)} U_t^{(s)} = \alpha^{(s)} + {\beta^{(s)}}' U^{(s)}$ is a linear function of the underlying T "factors" or "components" $U_1^{(s)},...,U_T^{(s)}$ (which are assumed to be independent, centered and normalized, with $T \leq b_s$) with unknown $\alpha^{(s)}$ and coefficients $\beta_t^{(s)}$. Thus, in row i of X, all data subvectors X_{i,B_s} are lying in the same T-dimensional subspace $H^{(s)}$ of R^{b_s} with coordinate vectors $U_{[i]}^{(s)} = (U_{i1}^{(s)},...,U_{iT}^{(s)})'$ (typically with $T = 1$

or 2). Typically this subspace will be different for different object groups A_r. Completing the corresponding index r in the previous notation, we obtain the *two-way subspace model*

$$X_{i,B_s} = \alpha_r^{(s)} + \beta_r^{(s)'} U_{[i]}^{(s)} \qquad \text{for } i \in A_r, r = 1,...,m, s = 1,...,n \qquad (14)$$

where the coordinate vectors $U_{[i]}^{(s)}$ are all supposed to be independent. Applying this model (under normal distribution assumptions) to the given data X, we obtain the following *two-way subspace clustering criterion*:

$$R(\mathscr{A},\mathscr{B},\alpha,\beta,u) := \sum_{r=1}^{m} \sum_{i \in A_r} \sum_{s=1}^{n} ||x_{i,B_s} - \alpha_r^{(s)} - \beta_r^{(s)'} u_{[i]}^{(s)}||^2 \to \min_{\mathscr{A},\mathscr{B},\alpha,\beta,u} \qquad (15)$$

which is to be minimized with respect to the parameters and the underlying (factor weighting) vectors $u_{[i]}^{(s)} = (u_{i1}^{(s)},...,u_{iT}^{(s)})' \in R^T$. Essentially this amounts to mn block-specific principal component analyses. After all, the component vectors $u_{[i]}^{(s)}$ can be displayed in R^T and then provide an idea about the configurations of the data within the data blocks $A_r \times B_s$. Similar models and algorithms are surveyed in Vichi (2012); quite generally they provide a remarkable reduction in data complexity in case of a large number J of variables that is reduced here to the dimension nT.

Finally we want to point to the fact that two-way clustering can also be seen in the context of (social) network analysis where we are given, in the simplest case, a data matrix that describes a binary relation among objects (rows) and properties (columns). The problem then consists in constructing blocks of objects (e.g., persons) with a similar behaviour with respect to the properties, and blocks of similarly related properties, all formulated in graphtheoretical terms. Suitable probabilistic and non-probabilistic models and methods are described, e.g., in the seminal publications by Holland and Leinhardt (1981); Anderson et al (1992); Wasserman and Faust (1994); Nowicki and Snijders (2001). Another approach is followed by Harris and Godehardt (1998); Godehardt and Jaworski (2003) and Godehardt et al (2010) who consider, to a given binary relation matrix, the corresponding “intersection graph" for objects and attributes, and analyze its properties in various probabilistic data models.

References

Anderson CJ, Wasserman S, Faust K (1992) Building stochastic blockmodels. Social Networks 14:137–161, DOI 10.1016/0378-8733(92)90017-2

Arabie P, Schleutermann S, Daws J, Hubert L (1988) Marketing applications of sequencing and partitioning of nonsymmetric and/or two-mode matrices. In: Gaul W, Schader M (eds) Data, Expert Knowledge and Decisions, Springer, Berlin, pp 215–224, DOI 10.1007/978-3-642-73489-2_18

Baier D, Gaul W, Schader M (1997) Two-mode overlapping clustering with applications to simultaneous benefit segmentation and market structuring. In: Klar R, Opitz O (eds) Classification and Knowledge Organization, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 557–566, DOI 10.1007/978-3-642-59051-1_58

Banerjee A, Merugu S, Dhillon IS, Ghosh J (2005) Clustering with Bregman divergences. The Journal of Machine Learning Research 6:1705–1749

Banerjee A, Dhillon IS, Ghosh J, Merugu S, Modha DS (2007) A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. The Journal of Machine Learning Research 8:1919–1986

Bocci L, Vicari D, Vichi M (2006) A mixture model for the classification of three-way proximity data. Computational Statistics & Data Analysis 50(7):1625–1654, DOI 10.1016/j.csda.2005.02.007

Bock HH (1980) Simultaneous clustering of objects and variables. In: Tomassone R, Amirchahy M, Néel D (eds) Analyse de données et informatique, INRIA, pp 187–203

Bock HH (1983) A clustering algorithm for choosing optimal classes for the chi-squared test. In: Contributed Papers, vol 2, Bull. 44th Session of the International Statistical Institute, pp 758–762

Bock HH (1992) A clustering technique for maximizing ϕ-divergence, noncentrality and discriminating power. In: Schader M (ed) Analyzing and Modeling Data and Knowledge, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 19–36, DOI 10.1007/978-3-642-46757-8_3

Bock HH (1996a) Probabilistic methods in cluster analysis. Computational Statistics and Data Analysis 23:5–28

Bock HH (1996b) Probability models and hypothesis testing in partitioning cluster analysis. In: Arabie P, Hubert LJ, De Soete G (eds) Clustering and classification, Studies in Classification, Data Analysis, and Knowledge Organization, World Scientific, Singapore, pp 377–453

Bock HH (2003) 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, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 143–154, DOI 10.1007/978-3-642-18991-3_17

Bock HH (2004) Convexity-based clustering criteria: Theory, algorithms, and applications in statistics. Statistical Methods and Applications 12(3):293–317, DOI 10.1007/s10260-003-0069-8

Castillo W, Trejos J (2002) Two-mode partitioning: Review of methods and application of tabu search. In: Jajuga K, Sokolowski A, Bock HH (eds) Classification, Clustering, and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 43–51, DOI 10.1007/978-3-642-56181-8_4

Celeux G, Diday E, Govaert G, Lechevallier Y, Ralambondrainy H (1989) Classification automatique des données : environnement statistique et informatique, Dunod, Paris, chap 2.6

Charrad M, Ben Ahmed M (2011) Simultaneous clustering: A survey. In: Kuznetsov SO, Mandal DP, Kundu MK, Pal SK (eds) Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science, vol 6744, Springer, Berlin, pp 370–375, DOI 10.1007/978-3-642-21786-9_60

Cheng Y, Church GM (2000) Biclustering of expression data. In: Proc. 8th International Conference on Intelligent Systems for Molecular Biology, vol 8, pp 93–103

Cho H, Dhillon IS (2008) Co-clustering of human cancer microarrays using Minimum Sum-Squared Residue co-clustering. IEEE/ACM Transactions on Computational Biology and Bioinformatics 5(3):385–400

Cho H, Dhillon IS, Guan Y, Sra S (2004) Minimum sumsquared residue co-clustering of gene expression data. In: Proc. 4th SIAM International Conference on Data Mining, pp 114–125

Dhillon IS, Mallela S, Modha DS (2003) Information-theoretic co-clustering. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, KDD '03, pp 89–98, DOI 10.1145/956750.956764

Gaul W, Schader M (1996) A new algorithm for two-mode clustering. In: Bock HH, Polasek W (eds) Data Analysis and Information Systems, Studies in

Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 15–23, DOI 10.1007/978-3-642-80098-6_2
Godehardt E, Jaworski J (2003) Two models of random intersection graphs for classification. In: Schwaiger M, Opitz O (eds) Exploratory Data Analysis in Empirical Research, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 67–81, DOI 10.1007/978-3-642-55721-7_8
Godehardt E, Jaworski J, Rybarczyk K (2010) Isolated vertices in random intersection graphs. In: Fink A, Lausen B, Seidel W, Ultsch A (eds) Advances in Data Analysis, Data Handling and Business Intelligence, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 135–145, DOI 10.1007/978-3-642-01044-6_12
Govaert G (1995) Simultaneous clustering of rows and columns. Control and Cybernetics 24(4):437–458
Govaert G, Nadif M (2003) Clustering with block mixure models. Pattern Recognition 36:463–473
Govaert G, Nadif M (2005) An EM algorithm for the block mixture model. Pattern Analysis and Machine Intelligence 27(4):643–647, DOI 10.1109/TPAMI.2005.69
Govaert G, Nadif M (2007) Block Bernoulli parsimonious clustering models. In: Brito P, Cucumel G, Bertrand P, de Carvalho F (eds) Selected Contributions in Data Analysis and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 203–212, DOI 10.1007/978-3-540-73560-1_19
Govaert G, Nadif M (2008) Block clustering with Bernoulli mixture models: Comparison of different approaches. Computational Statistics & Data Analysis 52(6):3233–3245, DOI 10.1016/j.csda.2007.09.007
Govaert G, Nadif M (2010) Latent block model for contingency table. Communications in Statistics - Theory and Methods 39(3):416–425, DOI 10.1080/03610920903140197
Govaert G, Nadif M (2013) Co-Clustering. Computing Engineering Series, Wiley, Chichester, UK
Hansohm J (2002) 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, Springer, Berlin, pp 87–93, DOI 10.1007/978-3-642-55991-4_9
Harris B, Godehardt E (1998) Probability models and limit theorems for random interval graphs with applications to cluster analysis. In: Balderjahn I, Mathar

R, Schader M (eds) Classification, Data Analysis, and Data Highways, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 54–61, DOI 10.1007/978-3-642-72087-1_6

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association 76(373):33–50, DOI 10.1080/01621459.1981.10477598

Kiers HAL, Vicari D, Vichi M (2005) Simultaneous classification and multidimensional scaling with external information. Psychometrika 70(3):433–460, DOI 10.1007/s11336-002-0998-4

Lazzeroni L, Owen A (2002) Plaid models for gene expression data. Statistica Sinica 12(1):61–86

Li J, Zha H (2006) Two-way Poisson mixture models for simultaneous document classification and word clustering. Computational Statistics & Data Analysis 50(1):163–180, DOI 10.1016/j.csda.2004.07.013

Li T (2005) A general model for clustering binary data. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, New York, KDD ’05, pp 188–197, DOI 10.1145/1081870.1081894

Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1(1):24–45, DOI 10.1109/TCBB.2004.2

Martella F, Vichi M (2012) Clustering microarray data using model-based double k-means. Journal of Applied Statistics 39(9):1853–1869, DOI 10.1080/02664763.2012.683172

Martella F, Alfò M, Vichi M (2008) Biclustering of gene expression data by an extension of mixtures of factor analyzers. The International Journal of Biostatistics 4(1)

Martella F, Alfò M, Vichi M (2011) Hierarchical mixture models for biclustering in microarray data. Statistical Modelling 11(6):489–505, DOI 10.1177/1471082X1001100602

Mirkin B, Arabie P, Hubert L (1995) Additive two-mode clustering: The error-variance approach revisited. Journal of Classification 12(2):243–263, DOI 10.1007/BF03040857

Nadif M, Govaert G (2010) Model-based co-clustering for continuous data. In: Draghici S, Khoshgoftaar TM, Palade V, Pedrycz W, Wani MA, Zhu X (eds) Proc. 2010 Ninth International Conference on Machine Learning and Applications (ICMLA’10), IEEE Computer Society, pp 175–180

Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic block-structures. Journal of the American Statistical Association 96(455):1077–1087, DOI 10.1198/016214501753208735

Rocci R, Vichi M (2008) Two-mode multi-partitioning. Computational Statistics & Data Analysis 52(4):1984–2003, DOI 10.1016/j.csda.2007.06.025

Schepers J, Hofmans J (2009) TwoMP: A MATLAB graphical user interface for two-mode partitioning. Behavior Research Methods 41(2):507–514, DOI 10.3758/BRM.41.2.507

Schepers J, van Mechelen I, Ceulemans E (2006) Three-mode partitioning. Computational Statistics & Data Analysis 51(3):1623–1642, DOI 10.1016/j.csda.2006.06.002

Schepers J, Ceulemans E, Van Mechelen I (2008) Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria. Journal of Classification 25(1):67–85, DOI 10.1007/s00357-008-9005-9

Schepers J, Bock HH, Van Mechelen I (2013) Maximal interaction two-mode clustering. Submitted

Shepard RN, Arabie P (1979) Additive clustering: Representation of similarities as combinations of discrete overlapping properties. Psychological Review 86(2):87–123, DOI 10.1037/0033-295X.86.2.87

Turner HL, Bailey TC, Krzanowski WJ, Hemingway CA (2005) Biclustering models for structured microarray data. IEEE/ACM Trans Computational Biology and Bioinformatics 2(4):316–329

Van Mechelen I, Bock HH, De Boeck P (2004) Two-mode clustering methods: A structured overview. Statistical Methods in Medical Research 13(5):363–394, DOI 10.1191/0962280204sm373ra

Van Rosmalen J, Groenen PJF, Trejos J, Castillo W (2009) Optimization strategies for two-mode partitioning. Journal of Classification 26(2):155–181, DOI 10.1007/s00357-009-9031-2

Vichi M (2001) Double k-means clustering for simultaneous classification of objects and variables. In: Borra S, Rocci R, Vichi M, Schader M (eds) Advances in Classification and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, pp 43–52, DOI 10.1007/978-3-642-59471-7_6

Vichi M (2012) Multimode clustering. In: Paper presented at the Symposium on Learning and Data Science (SLDS 2012), Firenze, Italy

Vichi M, Rocci R, Kiers HA (2007) Simultaneous component and clustering models for three-way data: Within and between approaches. Journal of Classification 24(1):71–98, DOI 10.1007/s00357-007-0006-x

Wasserman S, Faust K (1994) Social network analysis: Methods and applications, vol 8. Cambridge University Press, New York

Wilderjans TF, Depril D, Van Mechelen I (2013) Additive biclustering: A comparison of one new and two existing ALS algorithms. Journal of Classification 30(1):56–74, DOI 10.1007/s00357-013-9120-0