Developments in Data Analysis
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 12, Ljubljana: FDV, 1996

Probabilistic Models in Partitional Cluster Analysis

Hans H. Bock*

Abstract

Cluster analysis is designed for partitioning a set of objects into homogeneous classes by using observed data which carry information on the mutual similarity or dissimilarity of objects. Clustering methods are often defined in a heuristic or algorithmic way, emphasizing computational aspects and heuristic motivations. In contrast, this paper considers the clustering problem in a probabilistic framework and presents a survey on probabilistic models for partition-type clustering structures. It is shown how clustering criteria and grouping methods may be derived from these models in the case of vector-valued data, dissimilarity matrices and similarity relations.

1 The clustering problem and its underlying data

The ability to classify objects into homogeneous classes on the basis of their mutual similarities, dissimilarities or analogies is a basic element of human intelligence, an undispensible tool for the recognition of visual and conceptual structures, and an indispensable element for any abstract way of thinking. In the framework of statistics and data analysis, the classification problem occurs typically when large sets of objects are described by huge amounts of data which can never be analyzed without a preliminary step of information compression just by detecting or constructing a sufficiently small number of homogeneous classes of (similarly behaving) objects whose properties can be summarized by suitable class prototypes or class-specific feature combinations which provide an easy insight into, and a concise overview of, the full set of data. Or when it is conjectured that a given set of observations originates from several sources and seems to show some obvious heterogeneities: then the revelation of separate classes will reveal the hidden (possibly: causal) data structure and allow the development of class-specific strategies for solving substance-related questions, such as class-specific therapies for patients, group-specific publicity campaigns for attracting consumers, characterizing distinct types of social of psychological behaviour, distinguishing different types of soils or agricultural regions, locating single outlier cases etc. The classification of employees into different salary groups or the segmentation of the clients of an insurance company into types with their distinct risk structure provide examples with a more organizational motivation.

*Institute of Statistics, Technical University of Aachen, D-52056 Aachen, Germany

Clustering techniques are often considered as a part of exploratory statistics and many proposed clustering methods are just presenting a computational algorithm or some heuristic arguments, but proceed without any model assumptions or evaluation criteria. In contrast to such approaches this paper emphasizes a *model-based inferential approach* and presents a survey on some clustering or clustering-related methods which result from probability models for the underlying data.: We show how clustering strategies can be derived from these models by tools from classical statistics. Such an approach clarifies the conditions under which a proposed clustering method can be successful, and characterizes its performance.

Whilst we focus our presentation here on the description of probability models and the construction of related $k-$means like algorithms, it should be emphasized that the probabilistic approach extends to evaluation problems as well, e.g., when designing formal significance tests in order to check the existence of a 'clustering structure' or when defining models for 'purely random' (homogeneous) data constellations. This point of view is described, with many references in Bock (1974, 1985, 1989a, 1996a, 1996b), Perruchet (1983), Jain & Dubes (1988), Godehardt (1990) and Gordon (1994, 1995).

In the following, we consider a set $\mathcal{O} = \{1, ..., n\}$ of n objects $k \in \mathcal{O}$ described by data which are considered, in a probabilistic framework, as realizations of random variables. Then any inherent clustering (or non-clustering) structure for the objects must be characterized by the probability distribution of these variables. In this paper we will consider the following data types:

a) n *feature vectors* $x_1, ..., x_n$, each with p metric or qualitative components, describing the observed properties of the n objects. These data are realizations of n $p-$dimensional independent random vectors $X_1, ..., X_n$.

b) A *dissimilarity matrix* $(d_{kl})_{n \times n}$ where each entry d_{kl} characterizes the pairwise dissimilarity of the two objects $k, l \in \mathcal{O}$ (with $0 = d_{kk} \leq d_{kl} = d_{lk}$ for all k, l); they are realizations of $n(n-1)/2$ random dissimilarities D_{kl}, $k \neq l$ (with $0 \equiv D_{kk} \leq D_{kl} = D_{lk}$ for all k, l).

c) A *binary similarity relation* $(s_{kl})_{n \times n}$ where $s_{kl} = 1$ resp. $= 0$ indicates that the two objects k, l are considered to be 'similar' or not (with $s_{kk} = 1$ for all k), with corresponding binary random variables S_{kl}. These data are equivalent to a *similarity graph* G with n vertices (objects) and a link (edge) between two different vertices $k, l \in \mathcal{O}$ whenever $s_{kl} = 1$.

Whilst theoretically, any family of subsets of $\mathcal{O}$ may be considered as a 'classification' of $\mathcal{O}$, we will consider only *partitional classifications* here, i.e. a family $\mathcal{C} = (C_1, ..., C_m)$ with m non-empty non-overlapping subsets $C_i \subset \mathcal{O}$ with $\cup_{i=1}^{m} C_i = \mathcal{O}$ where m is a suitable (or specified) number of classes.

2 Partition-type models for random data vectors $\mathbf{X_1, ..., X_n}$

In this section we consider the case where the data are n random feature vectors $X_1, ..., X_n$ from R^p whose observed values are denoted by $x_1, ..., x_n$. We survey briefly six more or less common ways of defining a clustering structure for these data in terms of a probabilistic model. Thereby we must distinguish models which incorporate explicitly an m-partition of the objects from those which describe a 'clustering tendency' only.

2.1 The fixed-classification model

This model assumes that the distribution of the vectors is described by a known parametric family $f(\cdot;\vartheta)$ of distribution densities over R^p with a parameter $\vartheta \in R^q$. It supposes that there exists, for a fixed integer m, an unknown partition $\mathcal{C} = (C_1, ..., C_m)$ of $\mathcal{O}$ with m non-empty classes and m unknown class-specific parameters $\vartheta_1, ..., \vartheta_m$ compiled in the vector $\theta = (\vartheta_1, ..., \vartheta_m)$ such that:

$$X_k \sim f(\cdot;\vartheta_i) \qquad \text{for all } k \in C_i,\ i = 1, ..., m. \tag{2.1}$$

Assuming m to be known we can estimate the two 'parameters' $\mathcal{C}$ and θ by the maximum likelihood method which leads to the following *m.l. clustering criterion* (using the negative log likelihood):

$$g(\mathcal{C}, \theta) := \sum_{i=1}^{m} \sum_{k \in C_i} [-\log\ f(x_k;\vartheta_i)] \qquad \rightarrow \qquad \min_{\mathcal{C},\theta}. \tag{2.2}$$

Various special cases of this model will be considered in section 3. Here we mention only that an optimal pair $(\mathcal{C}, \theta)$ can be found or at least approximated by selecting an arbitrary (random or skilfully chosen) initial m-partition $\mathcal{C}^0$ and minimizing (2.2) with respect to $\mathcal{C}$ and θ in turn. The resulting well-known *k−means clustering algorithm* proceeds by iterating, for $t = 0, 1, 2, ...$, the following two steps:

(I) For the present partition $\mathcal{C}^t$ minimize (2.2) with respect to the unknown parameter θ. The solution is provided by the maximum likelihood estimate $\hat{\theta}^t := \theta(\mathcal{C}^t)$.

(II) For the obtained parameter vector $\theta = \theta^t$ minimize (2.2) with respect to the unknown m-partition $\mathcal{C}$. The solution is given by a so-called *minimum distance partition* or *maximum probability assignment* $\hat{\mathcal{C}}^{t+1} = \mathcal{C}(\theta^t)$ which comprizes the classes:

$$\hat{C}_i^{t+1} := \{k \in \mathcal{O} \mid f(x_k;\vartheta_i^t) = \max_{j=1,...,m} \{f(x_k;\vartheta_j^t)\}\ \}, \quad i = 1, ..., m, \tag{2.3}$$

where we must eventually adjust for ties in the 'boundaries' of the resulting classes, and care for possibly empty classes.

In fact, this algorithm yields a sequence C^0, θ^0, C^1, θ^1,... of successively improving partitions and parameter values which attains a stationary criterion value $g(C^t, \theta^t)$ after a finite number of iterations (*iterative minimum-distance clustering*, Bock 1974; *nuées dynamiques*, Schroeder 1976). Note that the convergence of the resulting sequence of *partitions* to a 'stationary' one has never been formally proved (theoretically, there might be a cycle of oscillating partitions in the end), but I know of no non-trivial example of such a behaviour, and in practice, a stationary partition is typically obtained after a few iterations. – Other approximate optimization methods (combinatorial enumeration, pairwise exchange strategies, dynamic programming, simulated annealing etc.) are described in Bock (1974), Späth (1985), Klein and Dubes (1989), Selim and Asultan (1991), Sun et al. (1994) and Hansen et al. (1994, 1996).

As an alternative to the maximum likelihood methods several authors have proposed a *Bayesian approach* which leads (under suitable prior assumptions and loss functions) to the optimization of a posterior risk (posterior probability) for the unknown m-partition C (see, e.g., Bock 1972, 1974, Binder 1978 and, for segmented prediction, Bernardo 1994).

2.2 The mixture model

Whilst a mixture density $f(x) = \sum_{i=1}^{m} \pi_i f(x; \vartheta_i)$ suggests intuitively m underlying classes or subpopulations Π_i described by class-specific densities $f(x; \vartheta_i)$, the usual mixture approach considers essentially n independent data vectors $X_1, ..., X_n$ all with the *same* (marginal) density $f(x)$ and focusses primarily on the estimation of the unknown parameters π_i and ϑ_i. In fact, such a model involves no explicit clustering of objects even if the corresponding (estimated) posterior probabilities $\hat{\pi}_i := \pi_i f(x_k; \hat{\vartheta}_i)/f(x_k)$ suggest a related 'fuzzy' classification.

As a more clustering-oriented alternative, we may consider, in addition to $X_1, ..., X_n$, m random binary class indicator vectors $I_1, ..., I_n$ where $I_k = e_i$ (the i–th unit vector in $\{0,1\}^m$) denotes that the k–th object (or X_k) originates from the i–th subpopulation Π_i such that I_k has a multinomial distribution $I_k \sim mult(1; \pi_1, ..., \pi_m)$ with probabilities $\pi_1, ..., \pi_m$ adding up to 1. These n indicators define a random (unobservable) partition C of $\mathcal{O}$ with classes $C_i = \{k \in \mathcal{O} | I_k = e_i\}$ which enters the joint likelihood function of the n i.i.d. pairs (I_k, X_k). When minimizing the minus log likelihood function $l(\pi, \theta; I_1, ..., I_n, x_1, ..., x_n)$, we minimize not only with respect to the unknown parameters π, θ, but also with respect to the 'missing values' $I_1, ..., I_n$ (equivalently: with respect to the induced partition C where the number of classes is bounded by m). Substituting the m.l. estimates $\hat{\pi}_i = |C_i|/n$ into $l(\cdot)$ yields finally the following partition-type clustering criterion:

$$\hat{g}(C,\ \theta) = \sum_{i=1}^{m} \sum_{k \in C_i} [-\log f(x_k; \vartheta_i)] - n \cdot \sum_{i=1}^{m} (|C_i|/n) \cdot \log(|C_i|/n) \rightarrow min_{C,\theta} \quad (2.4)$$

which adds an entropy term to the criterion (2.2) (Anderson 1985, Bock 1996a).

Mixtures are thoroughly investigated by Titterington, Smith & Makov (1985), Red-

ner & Walker (1984) and McLachlan & Basford (1988), the relationship to clustering and the determination of the class number m is fully discussed, e.g., in Windham (1987), McLachlan & Basford (1988), Windham & Cutler (1992, 1994), Furman & Lindsay (1994), Roeder (1994), Bozdogan (1994) and Bock (1996a).

2.3 Multimodality and high-density clusters

Any distribution density $f(x)$ can be characterized by its level sets $B(c) := \{x \in R^p | f(x) \geq c\}$ with c ranging between 0 and ∞. '*High-density clusters*' ('density-contour clusters') at a fixed level c are defined as the connected components $B_1(c), B_2(c), ...$ of the level set $B(c)$. They characterize, for a multimodal density f, the domains of local point aggregations when sampling from f, and when increasing the level c from 0 to ∞ they are successively reduced in size and split into subclasses in a pseudo-hierarchical way until they disappear at a sufficiently large level c (Bock 1974, chap. 28.c; Hartigan 1975, chap. 11.13, 1985).

Starting from n observed data points $x_1, ..., x_n$, suitable estimates $\hat{B}_i(c)$ for $B_i(c)$ can be obtained as the level sets of a (non-parametric or kernel-type) density estimate $\hat{f}$ of f. From these sets suitable object clusters $\hat{C}_i(c) := \hat{B}_i(c) \cap \{x_1, ..., x_n\}$ are easily constructed. It should be noted, however, that a visual display of high-density clusters or any easily interpretable formal description of their shape or of their boundaries is difficult for dimensions larger than two. – There exist many modifications of this basic clustering procedure, e.g. those using k-nearest neigbour distances and methods that start from discretized (grey-level) density values (e.g., for digitalized pictures) and incorporate morphological operations such as the dilatation and erosion of binary sets or the thinning and thickening of (boundary) functions, thus operations which are well-known from pattern recognition and image analysis (see Postaire 1993, Sbihi & Postaire 1995).

There were attempts to characterize the clustering tendency inherent in a density f respectively in the induced distribution P_f by real-valued functionals of f, e.g. by the *probability excess mass function* defined by

$$\begin{aligned} E(c) &:= \int [f(x) - c]^+ \, dx = \sum_{i=1}^{m} \int_{B_i(c)} [f(x) - c] \, dx \\ &\stackrel{*}{=} \sup_{(B_1,...,B_m)} \sum_{i=1}^{m} [(P_f(B_i) - c \cdot vol_p(B_i)] =: E^{(m)}(c). \end{aligned} \tag{2.5}$$

This index can be interpreted as the percentage of the density f which lies beyond the level c, or as the difference between the probability masses contained in the $B_i(c)$ under P_f and a uniform distribution, respectively. The equality $\stackrel{*}{=}$ holds for any m-modal density f if the supremum is taken over all sets of m disjoint connected subsets B_i of R^p. For the related theory and resulting clustering tests see Müller & Sawitzki (1991) and Sawitzki (1995).

2.4 Mode clusters

A closely related approach starts from the idea that the local maxima (modes) $\xi_1, \xi_2, \ldots$ of a (smooth) multimodal density f can be considered as the kernels of suitable cluster regions $D_1, D_2, \ldots$ in R^p where D_i is the set of all $x \in R^p$ which attain, after some hill-climbing relocation procedure (to be specified), the i-th mode ξ_i. Point clusters for a sample $x_1, \ldots, x_n$ are usually built up by a similar relocating process using a (smooth) density estimate $\hat{f}$ (Bock 1974, §28).

2.5 Clustered point processes

Spatial statistics provides a series of models for clustered point constellations $X_1, X_2, \ldots$ in an (often finite) domain $G \subset R^p$. Typical examples include the non-homogeneous *Poisson process* with a (multimodal) intensity function $\lambda(x)$, and the *Neyman-Scott process* where in a first stage parent Poisson points $Y_1, Y_2, \ldots$ are randomly located in G and in a second stage a random (Poisson distributed) number N_i of daughter points $X_{i1}, X_{i2}, \ldots, X_{iN_i}$ is located around Y_i (e.g., with a Gaussian distribution $\mathcal{N}_p(Y_i, \sigma^2 I_p)$ or with a uniform distribution in the ball $B(Y_i, r)$ for some radius $r > 0$). Statistical analysis concerns primarily the estimation of the incorporated parameters (λ, σ^2, r etc.; see Ripley 1981, Cressie 1991) and insofar the clustering tendency only (instead of locating single clusters).

2.6 Markovian models

A similar approach for modeling the clustering tendency of data points is provided by Markovian fields on R^p or on a lattice L of 'sites' (e.g., the pixels of a rectangular screen or of a discretized image). Considering, for brevity, this latter case for a finite rectangular $a \times b$ array of sites $L = \{k = (i,j) |\ i,j \text{ integer}, 1 \leq i \leq a, 1 \leq j \leq b\}$ in the two-dimensional lattice of integers in R^2, we may associate with each site $k \in L$ a random binary variable X_k which takes its value 1 (resp. 0) if the pixel k is coloured in black (resp. in white; similarly for more than two 'grey' levels). Since each pixel $k = (i,j)$ has four neigbours $k_1 = (i+1, j), k_2 = (i-1, j), k_3 = (i, j+1), k_4 = (i, j-1)$ (at the boundary of L there might be fewer of them), we may define 'clusters' as the connected sets of neighbouring black sites $k \in L$. The behaviour of these random clusters is described by the joint distribution of the ab random variables X_k which is modeled here by a *Markovian field*: For each $k \in L$, the conditional distribution of X_k given the other values $X_l, l \in L - \{k\}$, depends only on the colouring at the four neighbour sites k_1, k_2, k_3, k_4 which form the 'boundary' $\partial(k)$ of k. More specifically: $P(X_k = x | X_l = x_l, l \in L - \{k\}) = P(X_k = x | X_{k_1} = x_{k_1}, X_{k_2} = x_{k_2}, X_{k_3} = x_{k_3}, X_{k_4} = x_{k_4})$ for all $x, x_l, x_{k_\nu} \in \{0,1\}$ such that there may be $2^4 = 16$ different distributions for X_k which describe the tendency of X_k to be coloured like its four neigbours. Some of these distributions will be identical if we introduce horizontal or vertical homogeneity constraints, and the clustering or non-clustering tendency is controlled by suitable interaction parameters which are to be estimated from an observed 'picture' $\{x_k | k \in L\}$. – Markovian models are investigated, e.g., in Darroch et al. (1980), Wermuth & Lauritzen (1983), Cross & Jain (1983), Geman & Geman

(1984), Devijver & Dekesel (1988), Whittaker (1990) and Winkler (1994).

3 Some fixed-classification models for data vectors $X_1, ..., X_n$

The fixed classification model described in section 2. provides a very flexible tool for clustering purposes since suitable specifications of the density f (normal, double exponential etc.), of the class-specific parameters ϑ_i (central points or hyperplanes, variances, interactions etc.) and the inclusion of parameter constraints can cope with many special needs of practice and yield various interesting clustering methods. In the following we present a (by no means exhaustive) list of special cases together with a short description of the corresponding k-means algorithms. More details may be found in Bock (1974, 1987, 1996a, 1996b), Diday (1979), Späth (1985) and the following references.

3.1 The classical center-oriented normal distribution cases

(1) If we assume that all m hidden clusters have (approximately) the same spherical shape centered at a class-specific mean, we are led to the *normal distribution clustering model*:

$$X_k \sim \mathcal{N}_p(\mu_i, \sigma^2 I_p) \qquad \text{for all} \quad k \in C_i,\ i = 1, ..., m. \tag{3.1}$$

The maximum likelihood estimation of the unknown partition $\mathcal{C} = (C_1, ..., C_m)$ and the unknown parameter $\mu = (\mu_1, ..., \mu_m) \in R^{m \times p}$ leads to the clustering criterion:

$$g(\mathcal{C}, \mu) := \sum_{i=1}^{m} \sum_{k \in C_i} ||x_k - \mu_i||^2 \quad \longrightarrow \quad \min_{\mathcal{C}, \mu}, \tag{3.2}$$

and to its two essentially equivalent versions: the *variance criterion* (*sum-of-squares criterion*):

$$g(\mathcal{C}) := \sum_{i=1}^{m} \sum_{k \in C_i} ||x_k - \overline{x}_{C_i}||^2 \quad \longrightarrow \quad \min_{\mathcal{C}}, \tag{3.3}$$

and the *best-location criterion*:

$$\gamma(\mu) := \sum_{k=1}^{n} \min_{i=1,...,m} \{||x_k - \mu_i||^2\} \quad \longrightarrow \quad \min_{\mu}. \tag{3.4}$$

The corresponding classical k-means algorithm proceeds, for $t = 0, 1, 2, ...$, as follows:

(1.I) Calculate the class-specific means $\overline{x}_{C_1^t}, ...\overline{x}_{C_m^t}$ for the m classes of the partition $\mathcal{C}^t$;

(1.II) Take as the next partition $\mathcal{C}^{t+1}$ the minimum distance partition generated by the m class means resulting from (1.I) with classes:

$$\hat{C}_i^{t+1} := \{k \in \mathcal{O} \mid ||x_k - \overline{x}_{C_i^t}|| = \min_{j=1,...,m} \{||x_k - \overline{x}_{C_j^t}||\}\ \}, \qquad i = 1, ..., m. \tag{3.5}$$

(2) A closely related model results if we assume spherical clusters of varying diameters by assuming an $\mathcal{N}_p(\mu_i, \sigma_i^2 I_p)$ distribution in C_i with unknown variances σ_i^2 (Bock 1974, chap. 11). Proceeding one step further, we may allow for eventual correlations between the p coordinates of X_k. In fact, if we assume the *same* dependence structure in all m classes we obtain the model:

$$X_k \sim \mathcal{N}_p(\mu_i, \sigma^2 \cdot \Sigma) \qquad \text{for all} \quad k \in C_i,\ i = 1, ..., m. \tag{3.6}$$

with an unknown covariance matrix $\Sigma = Cov(X_k)$ such that the clusters correspond to m parallel ellipsoidal clouds of points scattered around the centers μ_i. The maximum likelihood approach yields here the *determinantal clustering criterion*:

$$|W(\mathcal{C})| \longrightarrow \min_{\mathcal{C}} \tag{3.7}$$

where the scatter matrix $W(\mathcal{C})$ is defined by:

$$W(\mathcal{C}) := \sum_{i=1}^{m} W(C_i) := \sum_{i=1}^{m} \sum_{k \in C_i} (x_k - \overline{x}_{C_i})(x_k - \overline{x}_{C_i})'. \tag{3.8}$$

Whilst (3.8) is the analogue of the variance criterion (3.3), it has an equivalent version which is the analogue of (3.2) (but not of (2.2)!):

$$|W(\mathcal{C}, \mu)| := |\sum_{i=1}^{m} \sum_{k \in C_i} (x_k - \mu_i)(x_k - \mu_i)'| \quad \longrightarrow \quad \min_{\mathcal{C}, \mu}. \tag{3.9}$$

The corresponding $k-$means algorithm proceeds essentially as before:

(2.I) Calculate the class-specific means $\overline{x}_{C_1^t}, ... \overline{x}_{C_m^t}$ for the m classes of the present partition $\mathcal{C}^t$, and estimate Σ by the corresponding scatter-based estimator $\hat{\Sigma}^t := W(\mathcal{C}^t)/n$.

(2.II) Determine, as the next partition $\mathcal{C}^{t+1}$, the minimum distance partition generated by the m class means $\overline{x}_{C_i}$ as in (3.5), but using the *Mahalanobis distance* induced by $\hat{\Sigma}^t$ or $W(\mathcal{C}^t)$ instead of the Euclidean one:

$$\hat{C}_i^{t+1} := \{k \in \mathcal{O} \mid ||x_k - \overline{x}_{C_i^t}||_{W(\mathcal{C}^t)^{-1}} = \min_{j=1,...,m} \{||x_k - \overline{x}_{C_j^t}||_{W(\mathcal{C}^t)^{-1}}\} \} \tag{3.10}$$

for $i = 1, ..., m$.

(3) Finally we may consider a normal model incorporating m *class-specific* dependence structures of the form:

$$X_k \sim \mathcal{N}_p(\mu_i, \Sigma_i) \qquad \text{for all} \quad k \in C_i,\ i = 1, ..., m \tag{3.11}$$

with m unknown covariance matrices Σ_i and point clusters of ellipsoidal shape, but with eventually different orientations and diameters. The corresponding likelihood clustering criterion is given by:

$$\prod_{i=1}^{m} |n_i^{-1} W(C_i)|^{n_i} \quad \longrightarrow \quad \min_{\mathcal{C}} \tag{3.12}$$

where $n_i := |C_i|$ denotes the size of the class C_i (with $n_1 + \cdots n_m = n$). This criterion is quite sensitive to near-singularity of any $W(C_i)$. The same holds for the alternative (non-equivalent) *modified-metric variance criterion*:

$$g(\mathcal{C}, Q, \mu) := \sum_{i=1}^{m} \sum_{k \in C_i} ||x_k - \mu_i||^2_{Q_i^{-1}} \quad \longrightarrow \quad \min_{\mathcal{C}, Q, \mu}, \tag{3.13}$$

where we minimize not only over $\mathcal{C}$ and $\mu = (\mu_1, ..., \mu_m)$, but also with respect to m positive definite matrices $Q_1, ..., Q_m$ (constrained by $|Q_i| = 1$) which determine the metrics which are used inside the classes C_i. The following *adaptive-distance clustering algorithm* is just the $k-$means algorithm for (3.13) (Diday & Govaert 1974, Späth 1985):

(3.I) Calculate the class-specific means $\overline{x}_{C_1^t}, ...\overline{x}_{C_m^t}$ for the m classes of the present partition $\mathcal{C}^t$;

(3.II) Determine the optimal metrics inside the classes by calculating the class-specific scatter matrices $\hat{Q}_i^t = W(C_i^t)/|W(C_i^t)|^{1/p}$, $i = 1, ..., m$;

(3.III) Determine, as the next partition $\mathcal{C}^{t+1}$, the minimum distance partition generated by the m class means $\overline{x}_{C_i^t}$ as in (3.10), but using in C_i the Mahalanobis distance induced by $\hat{Q}_i^t$:

$$\hat{C}_i^{t+1} := \{k \in \mathcal{O} \mid ||x_k - \overline{x}_{C_i^t}||_{(\hat{Q}_i^t)^{-1}} = \min_{j=1,...,m} \{||x_k - \overline{x}_{C_j^t}||_{(\hat{Q}_i^t)^{-1}}\} \} \tag{3.14}$$

for $i = 1, ..., m$.

It can be shown that this algorithm looks for a local minimum of the criterion:

$$g(\mathcal{C}) := \sum_{i=1}^{m} |W(C_i)|^{1/p} \longrightarrow \min_{\mathcal{C}}. \tag{3.15}$$

Note that the previous construction (3.14) must be amended by some further criteria which avoid the singularity of the resulting scatter matrices $W(\hat{C}_i^{t+1})$ in the next step (in particular: $|\hat{C}_i^{t+1}| > p + 1$). - Various aspects of clustering with non-Euclidean metrics are discussed by Marriott (1982) and Art et al. (1982).

3.2 Principal component clustering

Instead of characterizing each class C_i by its center point μ_i, it may be useful for some applications to characterize C_i by a class-specific *hyperplane* H_i of R^p of a given (low) dimension $s < p$, say, and to assume that all observation vectors X_k belonging to this same class are distributed near the hyperplane H_i. A corresponding normal distribution model assumes that there exists, in addition to the unknown m-partition $\mathcal{C} = (C_1, ..., C_m)$, a system of s-dimensional hyperplanes $\mathcal{H} = (H_1, ..., H_m)$ in R^p such that:

$$\begin{aligned} & X_k \sim \mathcal{N}_p(\mu_k, \sigma^2 I_p) \qquad \text{for all} \quad k = 1, ..., n \\ \text{with} \quad & \mu_k \in H_i \qquad \text{for all} \quad k \in C_i, \; i = 1, ..., m. \end{aligned} \tag{3.16}$$

Thus in each class the expectations $\mu_k = E[X_k]$ are all contained in the corresponding hyperplane H_i. Denoting by $P_H(x)$ the orthogonal projection of a point $x \in R^p$ onto a hyperplane H, the maximum likelihood method leads to the clustering criterion:

$$g(\mathcal{C}, \mathcal{H}, \mu) := \sum_{i=1}^{m} \sum_{k \in C_i} ||x_k - \mu_i||^2 \quad \longrightarrow \quad \min_{\mathcal{C},\mathcal{H},\mu \text{ under (3.16)}}, \tag{3.17}$$

and its two equivalent versions:

$$g(\mathcal{C}, \mathcal{H}) := \sum_{i=1}^{m} \sum_{k \in C_i} ||x_k - P_{H_i}(x_k)||^2 = \sum_{i=1}^{m} \sum_{k \in C_i} d(x_k, H_i) \quad \longrightarrow \quad \min_{\mathcal{C},\mathcal{H}} \tag{3.18}$$

and:

$$\gamma(\mathcal{H}) := \sum_{k=1}^{n} \min_{j=1,\ldots,m} d(x_k, H_i) \quad \longrightarrow \quad \min_{\mathcal{H}}. \tag{3.19}$$

where $d(x, H) := min_{y \in H}\{||x - y||^2\}$ is the squared orthogonal distance of a point $x \in R^p$ from a hyperplane H.

The corresponding *principal component clustering algorithm* has been described by Bock (1969, 1974, chap. 17, 1987) and Diday (1979, chap. 8) and proceeds as follows:

(4.I) For the given m-partitition C^t minimize (3.18) with respect to the hyperplanes H_i. The minimization of the i-th term in (3.18) is well-known from principal component analysis where it is shown that for the i-th class C_i^t the optimal hyperplane is given by $H_i^t = \overline{x}_{C_i^t} + [v_{i1}^t, ..., v_{is}^t]$, i.e. by the hyperplane which passes through the mean $\overline{x}_{C_i^t}$ and is spanned by the first s orthogonal eigenvectors $v_{i1}^t, ..., v_{is}^t \in R^p$ of the scatter matrix $W(C_i^t)$ (i.e. those which belong to the s largest eigenvalues).

(4.II) Determine, as the next partition C^{t+1}, the minimum-distance partition generated by the m hyperplanes H_i^t resulting from (4.I):

$$\hat{C}_i^{t+1} := \{k \in \mathcal{O} \mid d(x_k, H_i^t) = \min_{j=1,\ldots,m} \{d(x_k, H_j^t)\}\}, \quad i = 1, ..., m. \tag{3.20}$$

It is obvious how to generalize the clustering model (3.16) and the previous algorithm in order to allow for p dependent coordinates in X_k just by assuming an $\mathcal{N}_p(\mu_k, \Sigma)$ or an $\mathcal{N}_p(\mu_k, \Sigma_i)$ distribution in C_i. Another modification considers, in addition to the s *class-specific* dimensions (unit vectors) $v_{i1}, ..., v_{im}$ for H_i, some further *common* dimensions (unit vectors) $w_1, ..., w_r$ such that H_i has the form $H_i = a_i + [w_1, ..., w_r, v_{i1}, ..., v_{is}]$ with a dimension $r + s$ (Bock 1987). Generalizations of this type necessitate, however, a very large amount of data due to the numerous parameters to be estimated, and the interpretation and evaluation of the results is by no means an easy task.

3.3 Regression clustering

Instead of characterizing the classes by their principal component hyperplanes we may characterize them by regression hyperplanes as well. A suitable model needs, however, two types of data for each object $k \in \mathcal{O}$: A *deterministic* explanatory vector $z_k \in R^t$ and a *random* observation (target) vector $X_k \in R^p$. Then the data consist of the n pairs $(z_1, X_1), ..., (z_n, X_n)$ in R^{t+p}. An example is provided by n consumers where z_k describes some real-valued social or life-style characteristics of the k-th person while X_k contains the data on the products he buys or on his preference structure with respect to a basket of merchandises (for a discrete version see section 3.7 below). In these cases a p–dimensional characteristic hyperplane for the i-th class C_i may be put in the regression form $H_i = \{x = a_i + B_i z \mid z \in R^t\}$ with a fixed vector $a_i \in R^p$ and a $p \times t$ matrix B_i of unknown regression coefficients.

A corresponding fixed-classification model is provided by:

$$X_k \sim \mathcal{N}_p(\mu_k = a_i + B_i z_k, \sigma^2 I_p) \qquad \text{for all } k \in C_i,\ i = 1, ..., m \tag{3.21}$$

(Bock 1969, Charles 1977, Späth 1979, 1982). This model yields the same m.l. clustering criterion as in (3.18), but with the 'vertical' distance measure $d(x_k, H_i) := ||x_k - a_i - B_i z_k||^2$. The corresponding k–means *regression clustering algorithm* proceeds as follows:

(5.I) For each class C_i^t of the m-partitition C^t we calculate the regression hyperplane $H_i^t = \{x = \hat{a}_i^t + \hat{B}_i^t z \mid z \in R^t\}$ corresponding to the n_i^t observations in C_i^t and passing through the point $(\overline{z}_{C_i^t}, \overline{x}_{C_i^t})$ in R^{t+p}.

(5.II) We determine, as the next partition C^{t+1}, the minimum distance partition generated by the m hyperplanes H_i^t:

$$\hat{C}_i^{t+1} := \{k \in \mathcal{O} \mid d(x_k, H_i^t) = \min_{j=1,...,m} \{d(x_k, H_j^t)\}\ \}, \quad i = 1, ..., m. \tag{3.22}$$

Note that DeSarbo and Cron (1988) have proposed a mixture type model for regression clustering.

3.4 Projection pursuit clustering

For high-dimensional data $X_1, ..., X_n \in R^p$ it can be difficult to interpret the results of a clustering strategy in a direct way and we may therefore be interested in an optimal and suggestive visualization of the clustering structure in a low-dimensional (say: s-dimensional) representation of the data and clusters in R^s. The most straightforward solution is to display all observed data points $x_1, ..., x_n$ together with the constructed classification in the usual s-dimensional principal component plane, but this neglects totally the classification point-of-view. In contrast, the following clustering method combines the dimension reduction idea with the classification approach and will be particularly useful in situations where it is conjectured from the outset that in spite of a (high) dimension p of the data the main information provided by the

classes $C_1, ..., C_m$ is of a low-dimensional type.

The *common-hyperplane clustering model* assumes that there exists an unknown s-dimensional hyperplane $H = a + [v_1, ...v_s]$ in R^p such that all m class centers μ_i are elements of H:

$$X_k \sim \mathcal{N}_p(\mu_i, \sigma^2 I_p) \quad \text{for all } k \in C_i,\ i = 1, ..., m, \qquad (3.23)$$
$$\text{with} \quad \mu_i \in H \quad \text{for all} \quad i = 1, ..., m.$$

This model has been proposed and investigated by Bock (1987). With the estimates $\hat{a} = \overline{x}$ and $\hat{\mu}_i = P_H(\overline{x}_{C_i})$ it leads to the clustering criterion:

$$g(\mathcal{C}, H) := \sum_{i=1}^{m} n_i \cdot ||\overline{x}_{C_i} - P_H(\overline{x}_{C_i})||^2 + \sum_{i=1}^{m} \sum_{k \in C_i} ||x_k - \overline{x}_{C_i}||^2 \longrightarrow \min_{\mathcal{C}, H}. \qquad (3.24)$$

A suitable breakdown of this criterion yields, as a k-means strategy, the following *projection pursuit clustering algorithm* (Bock 1987):

(5.I) For the given m-partitition $\mathcal{C}^t$ minimize (3.24) with respect to the hyperplane H. The solution is provided by the hyperplane $H^t = \overline{x} + [v_1^t, ..., v_s^t]$ where $v_1^t, ..., v_s^t \in R^p$ are the s first eigenvectors of the *between-classes scatter matrix*

$$B(\mathcal{C}) := \sum_{i=1}^{m} n_i \cdot (\overline{x}_{C_i} - \overline{x})(\overline{x}_{C_i} - \overline{x})'$$

of $\mathcal{C}^t$, i.e. those which belong to the largest eigenvalues of this matrix.

(5.II) For the given hyperplane H^t, select the partition $\mathcal{C}^{t+1}$ which minimizes (3.24) with respect to $\mathcal{C}$. This is equivalent to minimizing the variance clustering criterion

$$g_{H^t}(\mathcal{C}) := \sum_{i=1}^{m} \sum_{k \in C_i} ||P_{H^t}(x_k) - P_{H^t}(\overline{x}_{C_i})||^2 \longrightarrow \min_{\mathcal{C}} \qquad (3.25)$$

for the n projected data points $P_{H^t}(x_k) \in H^t$. This minimization can be (approximately) performed, e.g., by applying the basic k-means algorithm (1.I), (1.II) to the projections $P_{H^t}(x_k)$.

It is interesting to note that this same algorithm has been obtained by Diday (1979, chap. 9: analyse typologique discriminante) when looking for a pair $(\mathcal{C}, H)$ which maximizes the variance of the projected points $P_H(x_k)$ *between* the m classes of $\mathcal{C}$, i.e. $\sum_{i=1}^{m} n_i ||P_H(\overline{x}_{C_i}) - P_H(\overline{x})||^2$.

3.5 Minimum volume clustering

Hardy & Rasson (1982) and Rasson et al. (1988) proposed and investigated a clustering model where each class C_i corresponds to a convex set S_i of R^p such that $S_1, ..., S_m$ are pairwise disjoint and have each a positive p-dimensional Lebesgue

measure $\lambda(S_i)$. Denoting by $\mathcal{R}(S)$ the uniform distribution in a convex set S, this model is given by:

$$X_k \sim \mathcal{R}(S_i) \qquad \text{for all } k \in C_i,\ i = 1,...,m. \tag{3.26}$$

with m unknown disjoint convex sets $S_1,...,S_m$. Since, for a sample $y_1,...,y_n$ from $\mathcal{R}(S_i)$, the maximum likelihood estimator for S_i is determined by the convex hull $\hat{S}_i = conv(y_1,...,y_n)$ of this sample (see Rasson 1979), the m.l. clustering criterion belonging to the model (3.25) is given by:

$$g_l(\mathcal{C}) := \sum_{i=1}^{m} |C_i| \cdot \log(vol(C_i)) \quad \longrightarrow \quad \min_{\mathcal{C}}. \tag{3.27}$$

where $vol(C_i) := \lambda(conv(\{x_k \mid k \in C_i\}))$ is a measure of the extent of the class C_i such that (3.27) can be interpreted as looking for convex hulls $\hat{S}_i$ as small as possible. A corresponding exchange clustering algorithm proceeds by successively checking for each data point x_k if it should be better transferred from its present class C_i to another one C_j. Denoting by $n_i = |C_i|$ the size of the class C_i, such a transfer will reduce the criterion (3.27) iff $n_i vol(C_i) - (n_i - 1)vol(C_i - \{k\}) > (n_j+1)vol(C_j+\{k\}) - n_j vol(C_j)$. Any such algorithm is computationally demanding since the multiple determination of convex hulls and of their volumes is very time-consuming.

3.6 Entropy clustering

Whilst in the previous sections we have considered quantitative data vectors, the fixed-classification model lends itself to the analysis of qualitative or nominal data as well. This will be illustrated here for the case where the p components of the observed data vectors $X_k = (X_{k1},...,X_{kp})'$ are of a qualitative (nominal) type. More specifically, suppose that the ν-th component $X_{k\nu}$ of X_k takes its values in a finite set of alternatives $\mathcal{X}_\nu = \{1,...,s_\nu\}$ such that X_k has its values in the cartesian product $\mathcal{X} := \prod_{\nu=1}^{p} \mathcal{X}_\nu$. In this case, the data can be summarized in a p-dimensional contingency table $\mathcal{N} = (n_y)_{y \in \mathcal{X}}$ where each of the $s := s_1 s_2 \cdots s_p$ cells $y = (y_1,...,y_p) \in \mathcal{X}$ contains the observed number $n_y \equiv n_{y_1 \cdots y_p}$ of observations k with $X_k = y$.

We want to partition the set $\mathcal{O}$ of objects into m classes C_i each characterized by a class-specific dependence structure between the components of X_k (cf. section 3.1). Such a dependence structure is usually modeled by a *loglinear model* involving a vector $\vartheta \in R^s$ of interaction (association) parameters and main effects. Due to space limitations we mention only that such a model involves typically a dummy binary vector $z(y) \in R^s$ which characterizes the location of each 'cell' $y \in \mathcal{X}$ in the contingency table. Then $P(X_k = y)$ is assumed to have the 'log-linear' form $P(X_k = y) = p(y,\vartheta) := \exp\{z'(y)\vartheta - \mu(\vartheta)\}$ for $y \in \mathcal{X}$ where the scalar product $z'(y)\vartheta$ picks from ϑ just the components needed for the cell y and $\mu(\vartheta) := \log(\sum_y \exp\{z'(y)\vartheta\})$ is a normalizing constant. – With this notation we consider the following probabilistic clustering model for the unknown m-partition $\mathcal{C}$:

$$P(X_k = y) = \exp\{z'(y)\vartheta_i - \mu(\vartheta_i)\} \text{ for } y \in \mathcal{X},\ k \in C_i,\ i = 1,...m. \tag{3.28}$$

For the data vectors $x_1, ..., x_n$ the m. l. clustering criterion is given by:

$$g(C, \theta) := \sum_{i=1}^{m} \{|C_i| \cdot \mu(\vartheta_i) - \sum_{k \in C_i} z'(x_k)\vartheta_i\} \quad \longrightarrow \quad \min_{C,\theta}. \tag{3.29}$$

A short calculation shows that this is equivalent to the following *entropy clustering criterion*:

$$g(C) := \sum_{i=1}^{m} |C_i| \cdot H(X, \hat{\vartheta}_i) \quad \longrightarrow \quad \min_{C}. \tag{3.30}$$

where $\hat{\vartheta}_i$ is the maximum likelihood estimate of the interaction vector in the class C_i and $H(X, \vartheta_i) := -\sum_{y \in \mathcal{X}} p(y, \vartheta_i) \cdot \log p(y, \vartheta_i) \geq 0$ is Shannon's entropy for the probability distribution $p(\cdot, \vartheta_i)$ in C_i. The resulting k-means algorithm is fully described in Bock (1986, 1994). Note that the present classification problem yields a decomposition of the global contingency table $\mathcal{N}$ into m tables $\mathcal{N}_i$ of the same size, but with distinct dependence structures. The method has also been proposed by Celeux & Govaert (1991).

3.7 Logistic regression clustering with entropy measures

In analogy to regression clustering for quantitative data (see section 3.3) we can formulate a logistic regression clustering model for nominal data as well. Here this will be exemplified for the binary case when we observe, for each object $k \in \mathcal{O}$, a random 0/1 target variable X_k which is assumed to be dependent on a non-random observed vector $z_k \in R^s$ which describes s explanatory variables (quantitative or qualitative) such that this dependence is given by a class-specific logistic model. More specifically, we assume that for our data $(z_1, X_1), ..., (z_n, X_n)$ there exists an unknown m-partition C and m class-specific parameters $\vartheta_1, ..., \vartheta_m \in R^s$ such that the distribution of X_k involves the linear combination $\beta_{ik} = z_k'\vartheta_i$ of the explanatory variables belonging to the k-th object and the logistic link function $p = h(\beta) = (1 + e^{-\beta})^{-1}$ in the following way:

$$P(X_k = 1) = \frac{1}{1 + e^{-z_k'\vartheta_i}} = h(\beta_{ik}) =: p_{ik} \quad \text{for } k \in C_i,\ i = 1, ...m \tag{3.31}$$

or, equivalently:

$$P(X_k = x) = p_{ik}^x (1 - p_{ik})^{1-x} = \exp\{x \cdot z_k'\vartheta_i - \alpha(z_k'\vartheta_i)\} \text{ for } x \in \{0, 1\}, \quad k \in C_i,\ i = 1, ...m. \tag{3.32}$$

with the function $\alpha(\beta) := \log(1 + e^{\beta}) = -\log(1 - p)$.

It can be shown that the m. l. clustering criterion (2.2) belonging to this model reduces, as an analogue to (3.3), to:

$$g(C) := \sum_{i=1}^{m} \sum_{k \in C_i} H(\hat{p}_{ik}, 1 - \hat{p}_{ik}) \quad \longrightarrow \quad \min_{C} \tag{3.33}$$

where $\hat{\vartheta}_i$ is the m.l. estimate of ϑ_i) in C_i which yields the estimated probabilities $\hat{p}_{ik} := h(z_k'\hat{\vartheta}_i)$, and $H(p, 1-p) := -[p \log p + (1-p)\log(1-p)]$ is the usual entropy function for a binary variable. The minimization of this entropy-based clustering criterion proceeds by the corresponding k-means algorithm:

(6.I) Calculate, for the given m-partitition $\mathcal{C}^t$, the m.l. estimates for θ_i by solving the likelihood equations:

$$\sum_{k \in C_i} x_k z_k = \sum_{k \in C_i} p_{ki} z_k \qquad i = 1, ..., m. \tag{3.34}$$

(6.II) Given the estimates $\hat{\vartheta}_1, ..., \hat{\vartheta}_m$, the maximum probability partition $\mathcal{C}^{t+1}$ is obtained by assigning each object $k \in \mathcal{O}$ to the class C_i^{t+1} for which:

$$\hat{p}_{ki} = h(z_k'\hat{\vartheta}_i) \to \max_i \quad \text{if } x_k = 1$$
$$\hat{p}_{ki} = h(z_k'\hat{\vartheta}_i) \to \min_i \quad \text{if } x_k = 0.$$

Details can be found in Bock (1986, 1994).

4 A probability model for dissimilarity data

Parametric and probabilistic clustering models for dissimilarity data have been rarely proposed in the literature. We describe here a fixed-classification model which has been proposed by Bock (1989b): Let us assume that the similarity relations between the n objects are determined by an observed $n \times n$ matrix (d_{kl}) of pairwise dissimilarities. The model starts with the idea that in a homogeneous or unstructured population all $\binom{n}{2}$ random nonnegative dissimilarities $\tilde{D}_{kl}$ with $k < l$ are independently distributed with the same (standardized) distribution, e.g., an exponential distribution $exp(1)$.

In contrast, the clustering model states that, for a fixed unknown m-partition $\mathcal{C} = (C_1, ..., C_m)$ of the objects, the observable dissimilarities D_{kl} with $k < l$ are distributed according to:

$$D_{kl} \quad \sim \quad \vartheta_{ij} \cdot \tilde{D}_{kl} \qquad \text{for all } k \in C_i,\ l \in C_j \tag{4.1}$$

with scaling factors $\vartheta_{ij} > 0$ which describe the reduction or increase of the standard dissimilarities *in* and *between* the classes, respectively (typically with side constraints $\vartheta_{ii} \leq \vartheta_{ij}$ for all i, j). They must be estimated, together with $\mathcal{C}$, from the observed matrix $(d_{kl})_{n \times n}$, e.g. by maximizing the likelihood. Note that the independence assumption is somewhat unrealistic due to the approximate transitivity property of real-case similarities and must be weakened for practical applications.

The method can be exemplified for the cited case of a standard exponential distribution where D_{kl} has the density $\vartheta_{ij}^{-1} \cdot \exp\{-d_{kl}/\vartheta_{ij}\}$ for $k \in C_i, l \in C_j$ (and $d_{kl} > 0$). This yields the (minus log) likelihood clustering criterion:

$$g(\mathcal{C}, \theta) := \sum_{1 \leq i \leq j \leq m} n_{ij}[\bar{d}_{ij}/\vartheta_{ij} + \log \vartheta_{ij}] \quad \longrightarrow \quad \min_{\mathcal{C}, \theta}. \tag{4.2}$$

where for $i \neq j$: $n_{ij} := |C_i| \cdot |C_j|$ and $\tilde{d}_{ij} := n_{ij}^{-1} \cdot \sum_{k \in C_i, l \in C_j} d_{kl}$ is the average dissimilarity between the classes C_i, C_j; and for $i = j$: $n_{ii} := \binom{n}{|C_i|}$ and $\tilde{d}_{ii} := n_{ii}^{-1} \cdot \sum_{k,l \in C_i, k<l} d_{kl}$.
Substituting here the (unconstrained) m.l. estimates $\hat{\vartheta}_{ij} = \tilde{d}_{ij}$ we obtain the very suggestive *log-distance clustering criterion*:

$$g(\mathcal{C}) := g(\mathcal{C}, \hat{\vartheta}) - \binom{n}{2} = \sum_{1 \leq i \leq j \leq m} n_{ij} \log \tilde{d}_{ij} \quad \longrightarrow \quad \min_{\mathcal{C}}. \tag{4.3}$$

5 Clustering models for random similarity relations

In this section we consider the case where, for any pair k, l of objects, only the two alternatives $s_{kl} = 1$ (i.e., the objects are rated to be 'similar') or $s_{kl} = 0$ (i.e., they are 'dissimilar') are possible. In our stochastic framework this leads to a random binary similarity relation $S = (S_{kl})$ on $\mathcal{O}$ (with $P(S_{kk} = 1, S_{kl} = S_{lk}$ for all $k, l) = 1$). S is equivalent to a random graph G with n vertices and a random number $N = \sum \sum_{k<l} S_{kl}$ of links $\overline{kl}$ with $S_{kl} = 1$. We mention three clustering models for this situation:

5.1 The fixed-classification model

This model postulates the existence of an unknown m-partition $\mathcal{C} = (C_1, ..., C_m)$ of $\mathcal{O}$ and of a symmetric matrix $p = (p_{ij})_{m \times m}$ of unknown class-specific linking probabilities $p_{ij} = p_{ji}$ (typically with $p_{ii} \geq p_{ij}$ for all i, j) such that all $\binom{n}{2}$ link indicators S_{kl} with $k < l$ are independently distributed with:

$$P(S_{kl} = 1) \; = \; p_{ij} \qquad \text{for all} \;\; k \in C_i, \; l \in C_j \tag{5.1}$$

(Bock 1989b). Applying the maximum likelihood method for estimating the unknown $\mathcal{C}$ and (p_{ij}) amounts to minimizing the clustering criterion:

$$g(\mathcal{C}, p) := - \sum_{1 \leq i \leq j \leq m} [N_{ij} \log p_{ij} + (n_{ij} - N_{ij}) \log(1 - p_{ij})] \; \rightarrow \min_{\mathcal{C}, p} \tag{5.2}$$

where for $i < j$ N_{ij} is the number of pairs $\{k, l\}$ with $k \in C_i, l \in C_j$ with a link $S_{kl} = 1$, and $n_{ij} = |C_i| \cdot |C_j|$ whilst for $i = j$ $n_{ii} = \binom{|C_i|}{2}$ denotes the number of different pairs $\{k, l\}$ with $k \in C_i, l \in C_j$, $k < l$ as in the previous section. Obviously $\hat{p}_{ij} := N_{ij}/n_{ij}$ is the m.l. estimate for p_{ij} if the side constraints are neglected.

5.2 An error perturbation model

This model describes the unknown partition $\mathcal{C}$ by an equivalence relation $\rho = (\rho_{kl})_{n \times n}$ with $\rho_{kl} = 1$ if and only if the objects $k, l \in \mathcal{O}$ belong to the same class of

$\mathcal{C}$. The model assumes that the indicators ρ_{kl} with $k < l$ are randomly perturbed in the way that $\rho_{kl} = 1$ is replaced by 0 with probability α, and $\rho_{kl} = 0$ is replaced by 1 with probability β, all perturbations being independent and symmetry being maintained. This yields an observable random symmetric reflexive relation $S = (S_{kl})_{n \times n}$ with $\binom{n}{2}$ independent entries and $P(S_{kl} = 1) = \rho_{kl}(1 - \alpha) + (1 - \rho_{kl})\beta$ for $k < l$. Suitable clustering methods have to estimate the unknown parameters α, β as well as the unknown partition C (including m) from the observed matrix S (Frank 1978). Note that this is a special case of the previous model 5.1 with $p_{ii} = 1 - \alpha$ and $p_{ij} = \beta$ for $i \neq j$.

5.3 Markov graphs for similarity relations:

Frank & Strauss (1986) have proposed a model for a random graph G, i.e. a joint distribution for the $\binom{n}{2}$ link indicators S_{kl}, which allows for some dependence between neighbouring links S_{kl}, S_{lt} sharing a common object l. More specifically, it is assumed that for each pair of object pairs $\{k, l\}, \{r, t\}$ the link indicators S_{kl}, S_{rt} are *conditionally* independent given all other indicators S_{uv}, provided that $\{k, l, r, t\}$ comprises 4 different objects (this excludes overlapping pairs $\{k, l\}$ and $\{l, t\}$ where conditional dependence is not excluded). It can be shown that the resulting marginal distribution of S is equivalent to a Markov field on a related graph Γ (whose vertices are the $\binom{n}{2}$ pairs of objects), and a classical theorem of Hammersley and Clifford states that the joint distribution of all S_{kl} with $k < l$ has, under some homogeneity and symmetry conditions, the form:

$$P(S = s) = const. \cdot \exp\{\alpha \cdot N_3(G) + \sum_{t=1}^{n-1} \beta_t \cdot M_t(G)\}. \tag{5.3}$$

where G is the graph corresponding to the given realization $s = (s_{kl})$ of S, $N_3(G)$ is the number of triads (complete subsets of size 3) in G, and $M_t(G)$ the number of $t-$stars (a $k \in \mathcal{O}$ linked with exactly t other objects) in G; $\alpha > 0$ and $\beta_t \in R$ are unknown model parameters for transitivity and clustering, respectively. The estimation of these parameters requires extensive analytical and computational efforts.

Similar models have been proposed in network analysis, e.g., by Holland & Leinhardt (1981), Bollobás (1985), Fienberg, Meyer & Wasserman (1985) and Wasserman & Anderson (1987). Banks & Carley (1994) give a survey and propose a probability model of the type $P(S = s) = c(\sigma) \cdot \exp\{\sigma \cdot d(s, s^*)\}$ for all s where s^* describes a fixed 'central' similarity graph (e.g., one implied by a partition C), $d(s, \tilde{s})$ is a measure of the deviation between two similarity relations $s, \tilde{s}$, and the dispersion or scaling parameter σ influences the normalizing constant $c(\sigma)$. The relation of these models (which were typically proposed in a sociological framework) to the definition and construction of clusters and classifications is not yet fully understood.

6 Conclusions

In this paper we have proposed a series of probability-based clustering models and derived suitable clustering criteria or clustering algorithms. These proposals could

be extended by statistical tests for clustering structures, evaluation methods for the resulting classifications, and similar models for hierarchical or tree-like classifications (see the references cited in the introduction). Whilst these methods are certainly useful for analyzing and assessing clustering tendencies, we want to conclude with the remark that it is evident that in practical situations typically no single probability model will completely fit the data. For example, real data sets will usually contain several types of clusters at the same time, so we must try to combine the positive results of several clustering and testing strategies in order to get an acceptable final classification.

References

[1] Art D., R. Gnanadesikan, and J.R. Kettenring (1982): Data-based metrics for cluster analysis. *Utilitas Mathematica*, **21A**, 75-99.

[2] Anderson J.J. (1985): Normal mixtures and the number of clusters problem. *Computational Statistics Quarterly*, **2**, 3-14.

[3] Banks D. and K. Carley (1994): Metric inference for social networks. *J. of Classification*, **11**, 121-149.

[4] Bernardo J.M. (1994): Optimizing prediction with hierarchical models: Bayesian clustering. In: P.R. Freeman, A.F.M. Smith (Eds.): *Aspects of uncertainty*. Wiley, New York, 1994, 67-76.

[5] Binder D.A. (1978): Bayesian cluster analysis. *Biometrika*, **65**, 31-38.

[6] Bock H.H. (1969): The equivalence of two extremal problems and its application to the iterative classification of multivariate data. Report of the Conference "Medizinische Statistik", Forschungsinstitut Oberwolfach, February 1969, 10pp.

[7] Bock H.H. (1972): Statistische Modelle und Bayes'sche Verfahren zur Bestimmung einer unbekannten Klassifikation normalverteilter zufälliger Vektoren. *Metrika*, **18**, 120-132.

[8] Bock H.H. (1974): *Automatische Klassifikation (Clusteranalyse)*. Vandenhoeck & Ruprecht, Göttingen, 480 pp.

[9] Bock H.H. (1977): On tests concerning the existence of a classification. In: *Proc. First Symposium on Data Analysis and Informatics, Versailles, 1977, Vol. II.* Institut de Recherche d'Informatique et d'Automatique (IRIA), Le Chesnay, 1977, 449-464.

[10] Bock H.H. (1985): On some significance tests in cluster analysis. *J. of Classification*, **2**, 77-108.

[11] Bock H.H. (1986): Loglinear models and entropy clustering methods for qualitative data. In: W. Gaul and M. Schader (Eds.): *Classification as a tool of research.* North Holland, Amsterdam, 1986, 19-26.

[12] Bock H.H. (1987): On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In: H. Bozdogan and A.K. Gupta (Eds.): *Multivariate statistical modeling and data analysis.* Reidel, Dordrecht, 1987, 17-34.

[13] Bock H.H. (Ed.) (1988): *Classification and related methods of data analysis.* North Holland, Amsterdam, 1988, 749 pp.

[14] Bock H.H. (1989a): Probabilistic aspects in cluster analysis. In: O. Opitz (Ed.): *Conceptual and numerical analysis of data.* Springer-Verlag, Heidelberg, 12-44.

[15] Bock H.H. (1989b): A probabilistic clustering model for graphs and similarity relations. Paper presented at the Fall Meeting 1989 of the Working Group 'Numerical Classification and Data Analysis' of the Gesellschaft für Klassifikation, Essen, November 1989.

[16] Bock H.H. (1994): Information and entropy in cluster analysis. In: H. Bozdogan et al. (Eds.): *Multivariate statistical modeling*, Vol. II. Proc. 1st US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Univ. of Tennessee, Knoxville, 1992. Kluwer, Dordrecht, 1994, 115-147.

[17] Bock H.H. (1996a): Probability models and hypotheses testing in partitioning cluster analysis. In: Ph. Arabie, G. De Soete and L. Hubert (Eds.): *Clustering and classification.* World Science Publishers, River Edge, NJ/USA, 1996, 377-453.

[18] Bock H.H. (1996b): Probabilistic models in partitional cluster analysis. Computational Statistics and Data Analysis. (Accepted for publication)

[19] Bock H.H. and P. Ihm (Eds.) (1991): *Classification, data analysis and knowledge organization.* Springer-Verlag, Heidelberg, 1991, 393 pp.

[20] Bock H.H., W. Lenski, and M.M. Richter (Eds.) (1994): *Information systems and data analysis: Prospects, foundations, applications.* Springer-Verlag, Heidelberg, 462 pp.

[21] Bock H.H. and W. Polasek (Eds.) (1996): *Data Analysis and Information Systems: Statistical and Conceptual Approaches.* Springer-Verlag, Heidelberg, 548 pp.

[22] Bollobás B. (1985): *Random graphs.* Academic Press, London.

[23] Bozdogan H. (1994): Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In: O. Opitz, B. Lausen, R. Klar (Eds.): *Information and classification: Concepts, methods and applications.* Springer-Verlag, Heidelberg, 1994, 40-54.

[24] Celeux G. and G. Govaert (1991): Clustering criteria for discrete data and latent class models. *J. of Classification*, **8**, 157-176.

[25] Charles Chr. (1977): *Regression typologique.* Rapport de Recherche no. 257. IRIA, Le Chesnay, 1977.

[26] Cressie N. (1991): *Statistics for spatial data.* Wiley, New York.

[27] Cross G.R. and A.K. Jain (1983): Markov random field texture models. *IEEE Trans. Pattern Analysis, Machine Intelligence*, **PAMII-5**, 24-39.

[28] Darroch J.N., S. Lauritzen, and T.P. Speed (1980): Markov-fields and log-linear models for contingency tables. *Ann. Statist.*, **8**, 522-539.

[29] DeSarbo W.S. and W.L. Cron (1988): A maximum likelihood methodology for clusterwise linear regression. *J. of Classification*, **5**, 249-282.

[30] Devijver P.A. and M.M. Dekesel (1988): Cluster analysis under Markovian dependence with applications to image segmentation. In: H.H. Bock (Ed.), 1988, 203-217.

[31] Diday E. and G. Govaert (1974): Classification avec distances adaptatives. *Comptes Rendues Acad. Sci. Paris*, **278**, série A, 993.

[32] Diday E. et al. (1979): *Optimisation en classification automatique*, Vol. I, II. Institut National de Recherche en Informatique et en Automatique (INRIA), Le Chesnay, France.

[33] Diday E., Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy (Eds.) (1994): *New approaches in classification and data analysis.* Proc. 4th Conference of the International Federation of Classification Societies (IFCS-93), Paris, 1993. Springer-Verlag, Heidelberg, 693 pp.

[34] Fienberg S.E., M.M. Meyer, and S.S. Wasserman (1985): Statistical analysis of multiple sociometric relations. *J. Amer. Statist. Assoc.*, **80**, 51-67.

[35] Frank O. (1978): Inferences concerning cluster structure. In: COMPSTAT 1978, Physica-Verlag, Würzburg, 259-265.

[36] Frank O. and D. Strauss (1986): Markov graphs. *J. Amer. Statist. Assoc.*, **81**, 832-842.

[37] Furman W.D. and B.G. Lindsay (1994): Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics and Data Analysis*, **17**, 473-492.

[38] Gaul W. and D. Pfeifer (Eds.) (1995): *From data to knowledge.* Proc. 18th Annual Conference of the Gesellschaft für Klassifikation, Oldenburg, 1994. Springer-Verlag, Heidelberg, 1995 (in press).

[39] Geman S. and D. Geman (1984): Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis, Machine Intelligence*, **PAMI-6**, 721-741.

[40] Godehardt E. (1990): *Graphs as structural models. The application of graphs and multigraphs in cluster analysis.* Friedrich Vieweg & Sohn, Braunschweig, 1990^2, 240 pp.

[41] Godehardt E. (1991): Multigraphs for the uncovering and testing of structures. In: H.H. Bock and P. Ihm (Eds.), 1991, 43-52.

[42] Godehardt E. and A. Horsch (1994): The testing of data structures with graph-theoretical models. In: H.H. Bock, W. Lenski, and M.M. Richter (Eds.), 1994, 226-241.

[43] Gordon A.D. (1994): Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, **18**, 561-581.

[44] Gordon A.D. (1995): Null models in cluster validation. In: W. Gaul and D. Pfeifer (Eds.), 1995 (in press).

[45] Hansen P., B. Jaumard, and E. Sanlaville (1994): Partitioning problems in cluster analysis: a review of mathematical programming approaches. In: E. Diday et al. (Eds.), 1994, 228-240.

[46] Hansen P. and B. Jaumard (1996): Computational methods in clustering from a mathematical programminf viewpoint. In: H.H. Bock and W. Polasek (Eds.), 1996, 24-40.

[47] Hardy A. (1994): An examination of procedures for determining the number of clusters in a data set. In: E. Diday et al. (Eds.), 1994, 178-185.

[48] Hardy A. and J.P. Rasson (1982): Une nouvelle approche des problèmes de classification automatique. *Statistique et Analyse des Données*, **7**, 41-56.

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

[50] Hartigan J.A. (1985): Statistical theory in clustering. *J. of Classification*, **2**, 63-76.

[51] Holland P.W. and S. Leinhardt (1981): An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.*, **76**, 33-65.

[52] Jain A.K. and R.C. Dubes (1988): *Algorithms for clustering data.* Prentice Hall, Englewood Cliffs, NJ.

[53] Klein R.W. and R.C. Dubes (1989): Experiments in projection and clustering by simulated annealing. *Pattern Recognition*, **22**, 213-220.

[54] Marriott F.H.C. (1982): Optimization methods of cluster analysis. *Biometrika*, **69**, 417-422.

[55] McLachlan G.J. and K.E. Basford (1988): *Mixture models: Inference and applications to clustering.* Marcel Dekker, New York and Basel.

[56] Müller D. W. and G. Sawitzki (1991): Excess mass estimates and tests for unimodality. *J. Amer. Statist. Assoc.*, **86**, 738-746.

[57] Perruchet C. (1983): Une analyse bibliographique des épreuves de classifiabilité en analyse des données. *Statistique et Analyse des Données*, **8**, 18-41.

[58] Postaire J.G., R.D. Zhang, and C. Botte-Lecocq (1993): Cluster analysis by binary morphology. *IEEE Trans. Pattern Anal. Machine Intell.*, **15**(2), 170-180.

[59] Rasson J.-P. (1979): Estimation de formes convexes du plan. *Statistique et Analyse des Données*, **46**, 31-46.

[60] Rasson J.-P., A. Hardy, and D. Weverbergh (1988): Point process, classification and data analysis. In: H.H. Bock (Ed.), 1988, 245-256.

[61] Redner R.A. and H.F. Walker (1984): Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195-239.

[62] Ripley B.D. (1981): *Spatial statistics.* Wiley, New York.

[63] Roeder K. (1994): A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.*, **89**, 487-495.

[64] Sawitzki G. (1995): The excess-mass approach and the analysis of multimodality. In: W. Gaul and D. Pfeifer (Eds.), 1995 (in preparation).

[65] Sbihi A. and J.-G. Postaire (1995): Mode extraction by multivalue morphology for cluster analysis. In: W. Gaul and D. Pfeifer (Eds.), 1994 (in preparation).

[66] Schroeder A. (1976): Analyse d'un mélange de distributions de probabilité de même type. *Revue de Statistique Appliquée*, **24**(1), 39-62.

[67] Selim S.Z. and K. Asultan (1991): A simulated annealing algorithm for the clustering problem. *Pattern Recognition*, **24**, 213-220.

[68] Silverman B.W. (1981): Using kernel density estimates to investigate multimodality. *J. Royal Statist. Soc.*, **B 43**, 97-99.

[69] Silverman B.W. and T. Brown (1978): Short distances, flat triangles and Poisson limits. *J. Applied Probability*, **15**, 816-826.

[70] Späth H. (1979): Clusterwise linear regression (algorithm 39). *Computing*, **22**, 367-373. Correction: **26**, 275.

[71] Späth H. (1982): A fast algorithm for clusterwise linear regression. *Computing*, **29**, 175-181.

[72] Späth H. (1985): *Cluster dissection and analysis. Theory, FORTRAN programs, examples.* Ellis Horwood, Chichester.

[73] Sun L.X., Y.L. Xie, X.H. Song, J.H. Wang, and R.Q. Yu (1994): Cluster analysis by simulated annealing. *Computers and Chemistry*, **18**, 103-108.

[74] Titterington D.M., A.F.M. Smith, and U.E. Makov (1985): *Statistical analysis of finite mixture distributions.* Wiley, New York.

[75] Wasserman S. and C. Anderson (1987): Stochastic a posteriori blockmodels: construction and assessment. *Social Networks*, **9**, 1-36.

[76] Wermuth N. and S. Lauritzen (1983): Graphical and recursive models for contingency tables. *Biometrika*, **70**, 537-552.

[77] Whittaker J. (1990): *Graphical models in applied multivariate statistics.* Wiley, Chichester.

[78] Windham M.P. (1987): Parameter modification for clustering criteria. *J. of Classification*, **4**, 191-214.

[79] Windham M.P. and A. Cutler (1992): Information ratios for validating mixture analyses. *J. Amer. Statist. Assoc.*, **87**, 1188-1192.

[80] Windham M.P. and A. Cutler (1994): Mixture analysis with noisy data. In: E. Diday et al. (Eds.), 1994, 155-160.

[81] Winkler G. (1994): *Image analysis: Markov fields and dynamic Monte Carlo methods.* Springer-Verlag, Heidelberg.