ELSEVIER Computational Statistics & Data Analysis 23 (1996) 5–28

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Probabilistic models in cluster analysis

Hans H. Bock

Institut für Statistik, Technische Hochschule Aachen, Wüllnerstr. 3, D-52056 Aachen, Germany

Abstract

This paper discusses cluster analysis in a probabilistic and inferential framework as opposed to more exploratory, heuristic or algorithmic approaches. It presents a broad survey on probabilistic models for partition-type, hierarchical and tree-like clustering structures and points to the relevant literature. It is shown how suitable clustering criteria or grouping methods may be derived from these models in the case of vector-valued data, dissimilarity matrices and similarity relations. In particular, we discuss hypothesis testing for homogeneity or for a grouping structure, the asymptotic distribution of test statistics, the use of random graph theory and combinatorial methods for simulating random dendrograms. Our presentation of hierarchies includes, e.g., Markovian branching processes and phylogenetic inference based on molecular sequence data.

Keywords: Probabilistic cluster analysis; Partition-type clustering; Hierarchical clustering models; Testing for a clustering structure; Phylogenetic inference

1. Cluster analysis: Data types and classification structures

Cluster analysis is designed to detect hidden 'groups' or 'clusters' in a set of objects which are described by numerical, linguistic or structural data such that the members of each cluster behave similarly to each other (with respect to the given data) and groups are hopefully well separated. Clustering techniques are often considered as a part of exploratory statistics, in particular if the used clustering algorithms are 'model-free' or only heuristically motivated. In contrast, this paper emphasizes an inferential approach and presents a brief survey on clustering and clustering-related methods which are based on probabilistic models. It shows how suitable clustering strategies can be derived from these models by tools from classical statistics, and analyzed in a probabilistic way. Such an approach clarifies the conditions under which

PII S0167-9473(96)018-7

a proposed clustering method can be successful, and characterizes its performance. Our presentation includes some formal test procedures for testing the existence of a 'clustering structure' as well as models for a 'purely random (homogeneous) data constellation'. Whilst this survey can point to many topics only very briefly, a more detailed account and additional references may be found in Bock (1974, 1985, 1989a, 1996a, b), Perruchet (1983), Jain and Dubes (1988), Godehardt (1990) and Gordon (1996a, b).

We consider a set $\mathcal{O} = \{1,\ldots,n\}$ of n objects $k = 1,\ldots,n$ 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 is characterized by the probability distribution of these variables. We will consider the following data types:

(a) *n feature vectors* $x_1,\ldots,x_n$, 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 $X_1,\ldots,X_n$;

(b) a *dissimilarity matrix* $(d_{kl})_{n\times n}$ with entries d_{kl} characterizing the dissimilarity of the 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 $D_{kk} \equiv 0$ for all k);

(c) a *binary similarity relation* $(s_{kl})_{n\times n}$ with $s_{kl} = 1$ resp. $= 0$ if the objects k,l are considered to be 'similar' or not (with $s_{kk} = 1$ for all k), with corresponding random Bernoulli 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 $s_{kl} = 1$.

In this paper we will consider two basic types of classification structures:

(1) *Partitions* $\mathcal{C} = \{C_1,\ldots,C_m\}$ of $\mathcal{O}$ with a suitable (or specified) number m of non-empty disjoint classes $C_1, C_2, \ldots \subseteq \mathcal{O}$;

(2) *Hierarchies* $\mathcal{H} = (A,B,\ldots)$ with nested classes $A,B,\ldots \subseteq \mathcal{O}$ (including all singletons as well as $\mathcal{O}$) such that $A \cap B \in \{A,B,\emptyset\}$ for all $A,B \in \mathcal{H}$, and *dendrograms* $(\mathcal{H},h)$ where $h \geq 0$ is an isotone heterogeneity index defined on the classes of $\mathcal{H}$.

Thus, we exclude here overlapping classifications (coverings) of $\mathcal{O}$ and fuzzy clustering concepts.

2. Partition-type models for data vectors $X_1,\ldots,X_n$

If the data are n random feature vectors $X_1,\ldots,X_n$, a probabilistic clustering model has been defined mainly in one of the five following ways (where we must distinguish models which incorporate explicitly an m-partition from those which describe a 'clustering tendency' only).

2.1. The fixed-classification model

This model assumes, for a fixed number m, an unknown m-partition $\mathcal{C} = (C_1,\ldots,C_m)$ of $\mathcal{O}$, m unknown class-specific parameters $\vartheta_1,\ldots,\vartheta_m$ compiled in the vector

$\theta = (\vartheta_1, \ldots, \vartheta_m)$ and a known parametric density family $f(\cdot; \vartheta)$ such that

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

If m is known we may estimate the two 'parameters' $\mathscr{C}$ and θ by the maximum likelihood method which leads to the following clustering criterion (using the negative log likelihood):

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

Minimizing with respect to $\mathscr{C}$ and θ in turn leads to the well-known k-means clustering algorithm which yields a sequence $\mathscr{C}^0$, θ^0, $\mathscr{C}^1$, $\theta^1, \ldots$ of successively improving partitions and parameter values (iterative minimum-distance clustering, nuées dynamiques; Bock 1974, Schroeder 1976). Other optimization methods (combinatorial, exchange, dynamic programming, etc.) are described in Bock (1974), Späth (1985), Hansen et al. (1994, 1996).

Fixed classification models provide a very flexible tool for clustering 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 special needs of practice and yield various interesting clustering methods:

- the *classical cases* which assume spherical or ellipsoidal normal distribution clusters with class-specific centers and lead, e.g., to the well-known *SSQ* (*trace*, *variance*) *criterion* $g(\mathscr{C}, \theta) = \sum_{i=1}^{m} \sum_{k \in C_i} ||x_k - \vartheta_i||^2$ and the determinantal criterion (Bock, 1974; Späth 1985);
- *principal component clustering* which assumes class-specific hyperplanes as in Bock (1969, 1974, 1987) and Diday (1973), possibly even with common *and* class-specific dimensions (Bock, 1987);
- *regression clustering* based on class-specific regression hyperplanes planes H_i for data (X_k, z_k) comprizing an explanatory vector z_k (Bock, 1969, 1987);
- *projection pursuit clustering* where all class centers $\vartheta_i \in R^p$ are located on an unknown low-dimensional hyperplane H of R^p (Bock, 1987, 1996a, b);
- *minimum-volume clustering* where classes correspond to convex sets in R^p (Rasson et al. 1988, Hardy 1994, Bock 1996c);
- *entropy clustering* for discrete data and assuming loglinear models for the X_k with class-specific interactions (Bock, 1986, 1994, 1996a, b, Céleux and Govaert, 1991);
- *binary regression clustering* (e.g. for credit scoring) yielding entropy criteria again (Bock, 1986, 1994, 1996a, b).

As an alternative to 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 $\mathscr{C}$ (see, e.g., Bock, 1972, 1974; Binder, 1978 and, for segmented prediction, Bernardo, 1994).

2.2. The mixture model and the random partition model

The usual marginal approach assumes the same mixture density $f(x) = \sum_{i=1}^{m} \pi_i f(x; \vartheta_i)$ for all vectors $X_1, \ldots, X_n$ with the purpose to estimate the unknown parameters π_i and ϑ_i: This model involves no explicit clustering. However, when considering, additionally, the random binary class indicator vectors $I_k \in \{0,1\}^m \sim Mult(1; \pi_1, \ldots, \pi_m)$ which define a random (unobservable) partition $\mathscr{C}$ of $\mathscr{O}$, the n i.i.d. pairs (I_k, X_k) yield a minus log likelihood function $l(\pi, \theta; I_1, \ldots, I_n, x_1, \ldots, x_n)$ which can be minimized with respect to π, θ and the missing values $I_1, \ldots, I_n$ (equivalently: with respect to the induced partition $\mathscr{C}$ where the number of classes is bounded by m). Substituting the m.l. estimates $\hat{\pi}_i = |C_i|/n$ into $l(\cdot)$ leads to the partition-type clustering criterion:

$$\hat{g}(\mathscr{C}, \theta) = \sum_{i=1}^{m} \sum_{k \in C_i} [-\log f(x_k; \vartheta_i)] - n \cdot \sum_{i=1}^{m} (|C_i|/n) \cdot \log(|C_i|/n) \to \min_{\mathscr{C}, \theta} \quad (2.3)$$

which adds an entropy term to the criterion (2.2) (Anderson, 1985). Mixtures are thoroughly investigated by Titterington, et al. (1985) and Redner and Walker (1984), the relationship to clustering and the determination of the class number m is fully discussed, e.g., in Windham (1987), McLachlan and Basford (1988), Windham and Cutler (1992, 1994), Furmann and Lindsay (1994), Roeder (1994), Bozdogan (1994) and Bock (1996a).

2.3. Multimodality and high-density (density-contour) clusters

Any density $f(x)$ is characterized by its level sets $B(c) := \{x \in R^p \mid f(x) \geq c\}$ for all $c > 0$. 'High-density clusters' at a fixed level c are defined as the connected components $B_1(c), B_2(c), \ldots$ of $B(c)$ which characterize, for a multimodal density f, the domains of point aggregations when sampling from f (Bock, 1974). Starting from n data points $\{x_1, \ldots, x_n\}$, corresponding estimates $\hat{B}_i(c)$ can be obtained from a (non-parametric or kernel-type) density estimate $\hat{f}$ of f from which suitable object clusters $\hat{C}_i(c) := \hat{B}_i(c) \cap \{x_1, \ldots, x_n\}$ are easily constructed. There exist many modifications, e.g., using k-nearest neighbour distances, and methods based on discretized (gray-level) density values which use morphological operations such as the dilatation and erosion of binary sets or the thinning and thickening of functions (well-known from pattern recognition and image analysis; see Postaire, 1993; Sbihi and Postaire, 1996).

The clustering tendency provided by f or by the induced distribution P_f can be characterized by the *probability excess mass function* given by

$$E(c) := \int [f(x) - c]^+ \, dx = \sum_{i=1}^{m} \int_{B_i(c)} [f(x) - c] \, dx$$

$$\stackrel{*}{=} \sup_{(B_1, \ldots, B_m)} \sum_{i=1}^{m} [(P_f(B_i) - c \cdot vol_p(B_i)] =: E^{(m)}(c), \quad (2.4)$$

i.e. 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 continuous density f and the supremum is taken over all sets of m disjoint connected subsets B_i of R^p (Müller and Sawitzki, 1991; Sawitzki, 1996).

2.4. Mode clusters

A closely related model starts from the idea that the local maxima $\xi_1, \xi_2, \ldots$ 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 ith 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, 1996a).

2.5. Point cluster processes

Spatial statistics provides a series of models for clustered point constellations $X_1, X_2, \ldots$ in a (often finite) domain $G \subset R^p$. Typical examples include the nonhomogeneous *Poisson process* with a (multimodal) intensity function $\lambda(x)$ and the *Neyman-Scott process* where parent points $Y_1, Y_2, \ldots$ are randomly located in G and a random (Poisson distributed) number N_i of daughter points $X_{i1}, X_{i2}, \ldots, X_{iN_i}$ is located near to Y_i (e.g., with a Gaussian distribution $N_p(Y_i, \sigma^2 I_p)$ or with a uniform distribution in the ball $K(Y_i, r)$ 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).

3. Partition-type probability models for dissimilarity data

Even if many clustering methods start from an $n \times n$ matrix (d_{kl}) of pairwise dissimilarities between objects, elaborated clustering models for this case are rarely found in the literature. The following fixed-classification approach has been proposed by Bock (1989b, 1996c): We start with the idea that in a homogeneous or unstructured population all $\binom{n}{2}$ random nonnegative dissimilarities $\tilde{D}_{kl}$ are independently distributed, all with the same (standardized) distribution, e.g., an exponential distribution exp(1). The clustering model states that, for a fixed unknown m-partition $\mathscr{C} = (C_1, \ldots, C_m)$ of the objects, the observed dissimilarities D_{kl} with $k < l$ are distributed according to:

$$D_{kl} \sim \vartheta_{ij} \cdot \tilde{D}_{kl} \quad \text{for all } k \in C_i,\ l \in C_j, \tag{3.1}$$

where the positive scaling factors ϑ_{ij} 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 $\mathscr{C}$, from the

observed data $(d_{kl})_{n\times n}$, e.g. by maximizing the likelihood. The independence assumption must be weakened for many practical applications.

Another model (Mountford, 1970) considers similarities S_{kl} instead of dissimilarities and proposes a normal distribution variance component model of the type $S_{kl} = \mu_{ij} + V_i + V_j + U_{kl}$ for all $k \in C_i$, $j \in C_j$ where μ_{ij} is the 'typical' dissimilarity between the classes C_i, C_j, $V_1, \ldots, V_m \sim N(0, \sigma^2)$ are class-specific deviations, and the $U_{kl} \sim N(0, \tau^2)$ denote random errors (all variables being independent). However, since Mountford considered the partition $\mathscr{C}$ to be known, he had no real clustering situation.

4. Partition-type clustering models for random similarity relations and random graphs

A random similarity relation $S = (S_{kl})$ on $\mathcal{O}$ (with $P(S_{kk} = 1) = 1$) tells us if two objects $k, l \in \mathcal{O}$ are considered to be 'similar' ($S_{kl} = 1$, a hit) or not ($S_{kl} = 0$, a failure). 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 $S_{kl} = 1$ as described in Section 1(c). Therefore, we may consider, occasionally, random graphs as well. We mention three clustering models here.

4.1. The fixed-classification model

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

$$P(S_{kl} = 1) = p_{ij} \quad \text{for all} \quad k \in C_i,\ l \in C_j. \tag{4.1}$$

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

$$g(\mathscr{C}, p) := -\sum_{1\leq i\leq j\leq m} [N_{ij} \log p_{ij} + (n_{ij} - N_{ij}) \log(1 - p_{ij})] \to \min_{\mathscr{C}, p}, \tag{4.2}$$

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

4.2. An error perturbation model

This model describes the unknown partition $\mathscr{C}$ 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 $\mathscr{C}$. 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 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 $\mathscr{C}$ (including m) from the observed matrix S (Frank, 1978). Note that this is a special case of the previous model 4.2 with $p_{ii} = 1-\alpha$ and $p_{ij} = \beta$ for $i \neq j$.

4.3. *Markov graphs for similarity relations*

Frank and 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 four different objects (this excludes overlapping pairs $\{k,l\}$ and $\{l,t\}$ where conditional dependence may exist). 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 the S_{kl} has, under some homogeneity and symmetry conditions, the form:

$$P(S = s) = \text{const.}\ \exp\left\{\alpha \cdot N_3(G) + \sum_{t=1}^{n-1} \beta_t \cdot M_t(G)\right\}, \tag{4.3}$$

where G is the graph corresponding to the considered 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 $k \in \mathscr{O}$ 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 and Leinhardt (1981), Bollobás (1985), Fienberg et al. (1985) and Wasserman and Anderson (1987). Banks and Carley (1994) give a survey and propose a probability model of the type $P(S = s) = c(\sigma)\exp\{\sigma \cdot d(s,s^*)\}$ for all s where s^* describes a 'central' similarity graph (e.g., implied by a standard partition $\mathscr{C}$), $d(s,\tilde{s})$ is a measure of the deviation between two similarity relations $s, \tilde{s}$, and the dispersion parameter σ influences the normalizing constant $c(\sigma)$.

5. Testing for homogeneity and for a clustering structure

Most clustering algorithms, including those which minimize a clustering criterion, provide the user always with a classification of objects – whether or not the data exhibit, in reality, a clear clustering structure and even if the calculated classes show

weak homogeneity or class separation properties. In order to prevent classificationists from pitfalls and wrong conclusions, it is therefore strongly recommended:

(a) to provide, before applying a clustering algorithm, some evidence that the data exhibit a clustering structure at all and are not, in the contrary, just a sample from a homogeneous universe;

(b) to assess, after having performed a cluster algorithm, the significance of the calculated classification or clusters such that finally only those classifications (clusters) are retained which are more marked than those resulting from 'random' data sets.

Problems of this type, together with the determination of a suitable number m of classes, concern a major part of recent theoretical and computational investigations in cluster analysis. Passing over a wealth of exploratory or interactive graphical tools, we will survey here a range of probability-based inferential methods (see also Bock, 1985, 1989a, 1996a, c; Gordon, 1994, 1996).

5.1. *Nearest-neighbour methods for testing for homogeneity*

Problems of type (a) are usually addressed by preliminary tests for homogeneity. The situation of 'homogenity' has been formalized either

- in the sense of a *uniform distribution* H_G of the data vectors $X_1,\ldots,X_n$ in a finite domain G of the space R^p, or
- in the sense of H_{unimod}, i.e. assuming an arbitrary *unimodal density* f for the X_k.

Test statistics for H_G are provided, e.g., by nearest-neighbour distances $D_k := \min_{l\neq k} ||X_l - X_k||$ between the n data points, the largest nearest-neigbour distance $T := \max_k\{D_k\}$ (possibly considering the boundary of G as well by using $T^* := \max_k\{\min\{D_k, ||X_k - \delta G||\}\}$), the radius R of the largest ball inside G without any data point in its interior, and modifications using the t-largest or s-smallest values. It appears that, for $n \to \infty$, the asymptotic distributions of T, T^* and R are all a rescaled Gumbel's extreme value distribution with distribution function $H(t) = \exp(e^{-t})$ such that percentage points can be easily approximated (at least for a known volume $|G|$). An asymptotic Smirnov type distribution results for the cited (and several weighted) modifications (see Henze, 1982; Dette and Henze, 1988; Janson, 1987; Bock, 1996a).

5.2. *Testing for multimodality*

A test for H_{unimod} versus bimodality H_2 or, more generally, for multimodality $H_{\leq m}$ with at most m modes versus $H_{>m}$ has been formulated by Silverman (1981; for $p = 1$) in terms of a kernel density estimator $\hat{f}$: H_{unimod} (respectively, $H_{\leq m}$) is rejected if the smallest critical window width h_{crit} for which $\hat{f}$ has 1 mode (respectively, m modes) is too large. Percentage points are obtained by bootstrap methods. Müller and Sawitzki (1991) and Sawitzki (1996) use an empirical version $E_n^{(m)}(c)$ of their excess mass statistics $E^{(m)}(c)$ (see (2.4)) and reject the hypothesis of m-modality $H_{\leq m}$

if, e.g., the maximum difference $D_{n,m}(c) := E_n^{(m+1)}(c) - E_n^{(m)}(c)$ (w.r. to $c > 0$) is too large. In the unidimensional case $p = 1$ the asymptotic distribution of $\max_c D_{n,m}(c)$ is described by a Brownian bridge, and for $m = 1$ the resulting test is basically equivalent to the DIP test proposed by Hartigan and Hartigan (1985).

Multivariate extensions of this latter test have been developed by Hartigan (1988; SPAN test), Hartigan and Mohanty (1992; RUNT test) and Rozál and Hartigan (1994; MAP test). They are all based on the edge lengths in the *minimum spanning tree* (MST) obtained for the Euclidean distances of the n data points $x_1, \ldots, x_n$ and insofar closely related to single linkage clustering. In particular, the largest edge length M_n (i.e. the level of the highest split in the single linkage dendrogram) has been investigated by Steele (1988), Steele and Tierney (1988) and Tabakis (1996). The latter paper derives asymptotic probability bounds for M_n which behaves as $[(\log n)/n]^{1/p}$ under a smooth, possibly multimodal density f (note that $M_n \geq T = \max_k\{D_k\}$). The empirical distribution of the $n - 1$ edge lengths in the MST has been considered by Pociecha and Sokolowski (1989) under multivariate normal and uniform distributions for the X_k.

5.3. *The max-F test and its generalizations*

Another range of tests concentrates on the previously mentioned problem (b) and checks the appropriateness of a calculated optimum m-partition $\mathscr{C}^* = \{C_1^*, \ldots, C_m^*\}$ by comparing a maximally (or minimally) attained clustering criterion value $k(\mathscr{C}^*) = \max_{\mathscr{C}} k(\mathscr{C})$ with a suitable percentage point $c = c(\alpha)$. For instance, the *max-F test* uses the ratio of the sum of squares between and in the classes of $\mathscr{C}$, i.e. the maximum value $k_{mn}^* = k_{mn}(\mathscr{C}^*)$ of:

$$k_{mn}(\mathscr{C}) = \left(\sum_{i=1}^{m} |C_i| \cdot ||\bar{x}_{C_i} - \bar{x}||^2 \right) \Big/ \left(\sum_{i=1}^{m} \sum_{k \in C_i} ||x_k - \bar{x}_{C_i}||^2 \right), \tag{5.1}$$

such that $\mathscr{C}^*$ will be the m-partition that minimizes the SSQ or variance criterion in the denominator. The asymptotic behaviour and distribution of k_{mn}^* and of the optimum class centers $\bar{x}_{C_i^*}$ for $n \to \infty$ has been theoretically investigated, e.g., by Bryant and Williamson (1978), Hartigan (1978; $p = 1$), Pollard (1982) and Bock (1985; $p \geq 1$); these asymptotics invoke an optimum m-partition $\mathscr{B}^* = (B_1^*, \ldots, B_m^*)$ of the Euclidean space R^p using a continuous version of the SSQ criterion. Whilst these theoretical results have been extended to generalized criteria as well (e.g., using generalized metrics instead of the Euclidean one), finite-sample distributions for k_{mn}^* and related test statistics (including the determinantal and Wilks type criteria; Lee, 1979) under 'homogeneity' must be computed by simulations (e.g., Jank 1996). A similar remark applies to the investigation of the power properties of these (and most) clustering tests under interesting clustering alternatives, a field which remains largely unexplored as yet.

5.4. Average similarities and U-statistics

It can be expected from an intuitive point of view that the average value $\bar{s}(\bar{d})$ of all $\binom{n}{2}$ similarities s_{kl} (dissimilarities d_{kl}) is larger (smaller) for a homogeneous population of the type H_{unimod} than for a clustered one. Therefore, average similarity or dissimilarity statistics have been occasionally proposed when testing for a clustering tendency. The investigation of such test statistics leads typically to the consideration of U-statistics (Bock, 1977, 1985; Silverman and Brown, 1978; Bhattacharya and Ghosh, 1992). Related proposals concerning a single cluster $C \in \mathcal{C}$ can be found in Gordon (1994, 1996) who uses, for pairs $\{k,l\},\{s,t\}$ of object pairs, binary distance comparison indicators $U_{kl,st} = 0, \frac{1}{2}, 1$ for $d_{kl} <, =, > d_{st}$ in order to define *local* and *global* validation indices for a given cluster C by $U_L := \sum_{k\in C} \sum_{l\in C-\{k\}} \sum_{t\in \bar{C}} U_{kl,kt}$ and $U_G := \sum_{k\in C, l\in C} \sum_{s\in C, l\in \bar{C}} U_{kl,st}$, respectively. These indices are used as test statistics for validating a constructed cluster C of objects (with significance points obtained by simulations).

5.5. Random graphs and multigraphs

In the case of a dissimilarity matrix (d_{kl}) a test for 'randomness' is often based on graph-theoretical concepts by considering, for a fixed, but arbitrary threshold $d \geq 0$, the threshold graph $G(d)$ with n vertices and a link $\overline{kl}$ for all pairs $k,l \in \mathcal{C}$, $k \neq l$, for which $d_{kl} \leq d$. $G(d)$ has exactly N edges if $d = d_{(N)}$, the N-smallest distance d_{kl} (if ties are neglected). Two (asymptotically equivalent) 'random graph' models are available for describing the 'pure randomness' of a graph:

(1) the *Bernoulli graph model* G_p which assumes $\binom{n}{2}$ independent link indicators $S_{kl} \sim Bin(1,p)$ all with the same linking probability $0 < p < 1$ (such that the number N of links in G_p has a $Bin(n,p)$ distribution), and

(2) the *combinatorial model* $\Gamma_{n,N}$ where a fixed number N of edges is randomly sampled over all $\binom{\binom{n}{2}}{N}$ possible selections.

For these models random graph theory provides a range of distributional results on the exact or asymptotic (for $n \to \infty$) distribution of clustering-related test statistics (Bollobás, 1985; Palmer, 1985). In the case of the threshold graph $G(d)$ with $d = d_{(N)}$, i.e. with N edges, we may consider, e.g.:

- the number N_{isol} of isolated vertices (one-element connected components $\{k\}$) of $G(d)$;
- the number N_{proper} of objects in proper (i.e. non-singleton) components of $G(d)$ such that $N_{\text{proper}} + N_{\text{isol}} = n$ holds trivially;
- the total number N_{comp} of connected components (single linkage clusters) at the level N;
- the size Z_{clique} of the largest clique in $G(d)$;
- the smallest integer $N = N_{\text{conn}}$ for which the threshold graph $G_{d(N)}$ is connected, etc.

The hypothesis of 'randomness' will be rejected in favour of a clustering structure (at the significance level α) if, e.g., N_{proper} or N_{comp} are smaller than their α-quantile.

The exact distribution of N_{proper} in $\Gamma_{n,N}$ is derived in Ling (1973, 1975) and Ling and Killough (1976). Many asymptotic results are surveyed or derived in Bollobás (1985), Nowicki (1988), Godehardt (1990, Chap. 5, 1992), Godehardt and Horsch (1994), Van Cutsem (1996) and Van Cutsem and Yeart (1996).

For instance, assuming $N = N(n) = \lfloor 0.5n(\log n + c + o(1))\rfloor$ for $n \to \infty$ with some constant $c > 0$, the graph $\Gamma_{n,N}$ consists asymptotically of one large component and several isolated objects, i.e. $P(N_{\text{proper}} = 1) \to 1$. Moreover, the random variables N_{isol} and $N_{\text{comp}} - 1$ are both asymptotically Poisson distributed $Po(\lambda)$ with expectation $\lambda = e^{-c}$, implying that $P(\Gamma_{n,N}$ is connected$) = P(N_{\text{comp}} = 1) \to \exp(e^{-c})$. Similar results relate to the number of vertices with a given degree m and to the number of isolated trees of size m in $\Gamma_{n,N}$ or G_{np}, but possibly for other choices of $N(n)$. Godehardt (1990, 1991) extends these results to the case of multigraphs $\Gamma_{n,N,t}$ which are the superposition of t layers of the type $\Gamma_{n,N}$, describing t different aspects of similarity for the same n objects. Matula (1970, 1972, 1976) shows that in the graph G_{np} the clique number Z_{clique} is asymptotically within $\lfloor z(n,p) - \varepsilon, z(n,p) + \varepsilon\rfloor$ for any $\varepsilon > 0$ where $z(n,p) := 2\log_{(1/p)} n - 2\log_{(1/p)} \log_{(1/p)} n + 2\log_{(1/p)}(\varepsilon/2) + 1$ and derives various finite-sample distributional properties.

The application of these graph models in cluster analysis is somewhat hindered by the fact that the assumption of (almost) independent links is unrealistic in many cases since natural similarity relations tend to be transitive (triangle inequality). Dissimilarity-based models with dependent links have been mentioned in Section 4 (see also Frank (1987) and the Euclidean incidence graphs in Godehardt and Horsch (1994)), but have not yet been fully developed in the 'homogeneous' or 'unimodal' case.

6. Probabilistic models for hierarchical and tree-like classifications

Hierarchical classifications are broadly used in applications in order to get a stratified structure of classes at various heterogeneity levels and to visualize the mutual similarities between classes in a two-dimensional display. A hierarchical classification is usually constructed in the form of a dendrogram $(\mathscr{H}, h)$, where $\mathscr{H}$ is a hierarchy of sets and h a numerical index on $\mathscr{H}$ such that $h(A)$ measures the heterogeneity of a class $A \in \mathscr{H}$ of objects in terms of the data. It is well-known that a dendrogram can be equivalently described by the ultrametric dissimilarity matrix $\delta = (\delta_{kl})_{n\times n}$ where $\delta_{kl} = \min\{h(A) \mid A \in \mathscr{H},\ k,l \in A\}$ is the heterogeneity $h(A)$ of the smallest class $A \in \mathscr{H}$ that contains both objects k and l. This dissimilarity measure fulfills the ultrametric inequality $\delta_{kl} \le \max\{\delta_{kj}, \delta_{jl}\}$ for all $j,k,l \in \mathcal{O}$ (which implies the triangle inequality), and the classes of $\mathscr{H}$ are just the δ-balls $A = B(k,d) := \{l \in \mathcal{O} \mid \delta_{kl} \le d\}$ (for $d \ge 0$ and $k \in \mathcal{O}$) while $h(A)$ is the δ-diameter of A.

Even if hierarchical clustering algorithms are often applied to data points $x_1, \ldots, x_n$ in the Euclidean space, it seems to be difficult to design a general, intuitive and spatial idea of a 'natural' *hierarchical* classification in R^p. In fact, hierarchical classifications suggest more an underlying evolutionary or branching process in time or space and the related probabilistic models are therefore often defined in terms of stochastic

processes. In the following sections we will review models based on dendrograms, additive trees, Markov processes and combinatorial considerations.

6.1. The additive error model for dissimilarity data

The *additive error model* proposed by Degens (1983) assumes that the observed dissimilarity matrix (d_{kl}) reflects an underlying unknown dendrogram structure $(\mathscr{H}, h)$ up to some random error. More specifically, if $\delta = (\delta_{kl})$ is the ultrametric which characterizes $(\mathscr{H}, h)$, the random (observed) dissimilarities

$$D_{kl} = \delta_{kl} + U_{kl} \quad \text{for } k, l \in \mathcal{O}, k \neq l \tag{6.1}$$

are obtained from δ_{kl} by independent additive error terms U_{kl} all with the same distribution density $\psi(\cdot)$ on R (e.g., a normal density). An estimate for δ resp. for $(\mathscr{H}, h)$ is then obtained by maximizing the likelihood $L := \prod\prod_{k<l} \psi(d_{kl} - \delta_{kl})$ over all ultrametrics δ and all unknown parameters in ψ (e.g., by combinatorial or heuristic algorithms).

In particular, if we assume a normal distribution $N(0, \sigma^2)$ for ψ, the maximization of L amounts to minimizing the SSQ error criterion $\sum\sum_{k<l}(d_{kl} - \delta_{kl})^2$ over δ (for combinatorial and penalty function minimization methods see, e.g., De Soete et al. (1987), De Soete (1988) and Sriram and Lewis (1993)) and it turns out that the level $h(A)$ of any class A of the resulting optimum dendrogram $(\mathscr{H}, h)$ is necessarily the average of the observed distances of its direct predecessors $B, C \in \mathscr{H}$ with $B + C = A$ (inducing a generalized average linkage algorithm). Analogous results have been obtained for other situations as well, e.g., for ψ a two-sided exponential distribution (generalized median procedure), for $U_{kl} \geq 0$ with a decreasing ψ (single linkage method), or $U_{kl} \leq 0$ with an increasing ψ (modified complete linkage method). These results were obtained by Degens (1983, 1985, 1988) together with several generalizations which refer, e.g., to the analysis of genetic distance data (d_{kl}) obtained from DNA–DNA hybridization experiments. Since these experiments can be repeated several times for each pair of species (objects) $k, l \in \mathcal{O}$ it is possible to design a similar Gaussian additive error model with *pair-specific variances* $\sigma_{kl}^2 = \mathrm{Var}(D_{kl}) = \mathrm{Var}(U_{kl})$ which can be estimated from the replicated measurements of D_{kl}. This approach leads to hierarchical weighted average linkage algorithms and yields theoretical consistency results.

6.2. Variance component models for genetic distance data

Replicated genetic distance data d_{kl} have been typically used for reconstructing the evolutionary tree of n given species in the form of a dendrogram. Since the evolution of these species proceeds partially in parallel (or identical) streams in the past, the independence assumption from the previous section will be unrealistic here. Therefore, Lausen and Degens (1986) and Degens et al. (1988) have proposed and investigated several *variance component models* which take into account various causes of variability of the measurements and from the evolutionary process

which leads to a nontrivial dependence structure for the dissimilarities D_{kl}. A typical example is given by

$$D_{klv} = \delta_{kl} + E_{kl} + \sum_{j \in P(k,l)} L_j + U_{klv} \quad \text{for } k,l \in \mathcal{C},\ k \neq l,\ v = 1,\ldots,n_{kl} \tag{6.2}$$

where $E_{kl} \sim N(0,\sigma_e^2)$ describes a pair-specific variation, $P(k,l)$ is the set of edges j in the path joining the object k with the object l (in the dendrogram belonging to (δ_{kl})), L_j the random evolutionary fluctuation existing along these edges, and $U_{klv} \sim N(0,\sigma_U^2)$ is the measurement error for the replications $v = 1,\ldots,n_{kl}$ of D_{kl}. The accuracy and stability of the resulting phylogeny and the induced hierarchy can be checked by using weighted three- or four-objects estimators for the variances (Wolf and Degens, 1991).

6.3. *Phylogenies and evolutionary Markov models for molecular sequences*

Whilst phylogenetic trees and dendrograms have been constructed from morphological data of species and from genetic distance matrices since a long time, the advent of fast and precise sequencing methods in molecular biology has revolutionized this field: Phylogenetic inference of n species $k = 1,\ldots,n$ is nowadays primarily based on the analysis of the corresponding nucleic acid sequences $x_k = (x_{k1},\ldots,x_{kp})$ with p contiguous sites $j = 1,\ldots,p$ and components x_{kj} taken from an alphabet $\mathcal{A} = \{A,G,C,T\}$ with $s = 4$ 'letters' which represent the four nucleotides, bases or 'states' adenine, guanine, cytosine, and thymidine (for DNA strains); thus there are 4^p different sequences. Similar data are available, e.g., for RNA ($s = 4$) or proteins ($s = 20$). Assuming that the underlying phylogeny has evolved in time t by random mutations (substitutions) of the bases in single sites, probability models for the resulting branching process have been formulated in terms of a (homogeneous) continuous-time Markov process for each site, and the phylogenetic tree is estimated by maximum likelihood, as well as its branch lengths and other model parameters (substitution rates).

A simple model (Felsenstein, 1981; Bishop and Friday, 1985) assumes that the mutation times for any single site j form a Poisson process with rate λ and that, if a mutation takes place from a state ν, it flips to the state μ with a probability π_μ (e.g., $\pi_\mu = 1/4$ for all $\mu \in \mathcal{A}$). This yields the transition probability $P_{\nu\mu}(t) = \pi_\mu(1 - e^{-\lambda t}) + \delta^*_{\nu\mu} e^{-\lambda t}$ from ν to μ in a time period t (where $\delta^*_{\nu\mu}$ denotes Kronecker's delta). More general cases are based on a matrix $R = (r_{\nu\mu})$ of substitution rates $r_{\nu\mu}$ (with zero row sums) such that the matrix of transition probabilities is given by $P(t) = (P_{\nu\mu}(t)) = e^{Rt}$, and the probabilities π_μ correspond to the stationary distribution of the induced homogeneous Markov process.

The relationship between the n species is described by an (unrooted unweighted) tree T with n leaves $k = 1,\ldots,n$ characterizing the n given species and a number m of interior vertices $k = n+1,\ldots,n+m$ representing m unobservable intermediate species in the past, together with branch lengths $t_1,\ldots,t_M$ which represent the time difference between two change points (for binary trees: $m = n-2$ and $M = n-3$). In order to write down the likelihood function (6.4) below, it will be appropriate to

specify one arbitrary interior node k^* of T as a root ($k^* = m+n$, say, a hypothetical ancestor) such that T becomes a directed graph T^* and its M branches can be written in the form $e_l = (a_l, b_l)$ where the node a_l is the direct ancestor of the node b_l in T^* and the numeration is such that $e_1, \ldots, e_n$ end in the observed leaves of the tree whilst each of the remaining edges $e_{n+1}, \ldots, e_M$ connects two interior points of T^*.

Our data consist of n sequences $x_k = (x_{k1}, \ldots, x_{kp}) \in \mathscr{A}^p$ observed for p neighbouring sites of n molecular strains (note that we pass over all problems of optimum alignment here). Additionally, we have to consider the (unobservable) base sequences $y_k = (y_{k1}, \ldots, y_{kp}) \in \mathscr{A}^p$ characterizing the putative intermediate species (interior nodes) $k = n+1, \ldots, n+m$ in the mutation process. For ease of notation we consider a 'homology' model where all sites $j = 1, \ldots, p$ evolve independently and identically in time and the mutation rates λ_l are identical for all branches e_l of the tree. Under these assumptions, the likelihood of our sample $x_1, \ldots, x_n$ is given by

$$L(T, \vartheta; x_1, \ldots, x_n) = \prod_{j=1}^{p} L_j(T, \vartheta; x_{1j}, \ldots, x_{nj}) \tag{6.3}$$

with factors L_j given by

$$L_j(T, \vartheta; x_{1j}, \ldots, x_{nj}) = \sum_{y_{n+1,j}, \ldots, y_{n+m,j}} \pi_{y_{m+n,j}} \cdot \prod_{l=1}^{n} P_{y_{a_l,j} x_{b_l,j}}(t_l; \vartheta) \cdot \prod_{l=n+1}^{M} P_{y_{a_l,j} y_{b_l,j}}(t_l; \vartheta) \tag{6.4}$$

and transition functions $P_{\nu\mu}(t; \vartheta)$ as specified before. Here ϑ contains all unknown parameter values of the Markov process, i.e., $\lambda_1, \ldots, \lambda_M$, $t_1, \ldots, t_M$ (in fact, only the products $\lambda_l t_l$ occur here), $p_{\nu\mu}$ and $r_{\nu\mu}$.

The usual procedure for estimating the phylogenetic tree T and the unknown parameter ϑ is provided by the maximum likelihood method which includes, in particular, the maximization of L over all tree topologies T. Excellent reviews of these methods were given by Felsenstein (1983a, 1988) and Goldman (1990), and many more or less general models have been proposed by Kimura (1980), Felsenstein (1981; all λ_q alike), Hasegawa et al. (1985), Cavender and Felsenstein (1987), Barry and Hartigan (1987; 12 parameters per branch), Lausen (1989, 1991), Navidi et al. (1993) and Schöniger et al. (1994). Various maximization or tree construction algorithms were described by Hendy and Penny (1982), Felsenstein (1981, 1990), Guénoche (1993a, 1993b), Barry and Hartigan (1987) and Navidi et al. (1993). A uniqueness theorem for the likelihood solution was proved by Fukami and Tateno (1989). A Bayesian approach has been followed by Felsenstein (1984) and Kishino and Hasegawa (1989).

Another likelihood approach is related to classical *parsimony methods* where the tree T and the unobservable ancestor sequences $y_1, \ldots, y_M$ are chosen such that, e.g., the total number N_{tot} of mutations along all branches is minimized (Wagner trees). This approach considers $y_1, \ldots, y_M$ as incidental parameters and maximizes the corresponding likelihood $L(T, \vartheta, y_1, \ldots, y_M; x_1, \ldots, x_n)$ with respect to these sequences as well. For example, in the case $\mathscr{A} = \{0, 1\}$ with $s = 2$ states, let us denote by

$\vartheta = 1 - \pi \in [0,1]$ the probability of a change $0 \to 1$ or $1 \to 0$ along a branch of T (in the previous model this corresponds to constant average times $\lambda_l t_l = c$, say, and equal transition probabilities $1 - \pi = P_{01}(c) = P_{10}(c)$). Then the likelihood has the form:

$$L(T, \pi, y_1, \ldots, y_M; x_1, \ldots, x_n) = \pi^{\sum_j (N_{00}^j + N_{11}^j)} \cdot (1-\pi)^{\sum_j (N_{01}^j + N_{10}^j)}, \tag{6.5}$$

where $N_{v\mu}^j$ denotes the number of branches e_l in T with character values (v, μ) at their end nodes (in site j). Since both sums add to the constant $Mp = (2n-3)p$ for binary trees, the maximization of L with respect to T and $y_1, \ldots, y_m$ amounts to minimizing $N_{\text{tot}} = \sum_j (N_{01}^j + N_{10}^j)$, the *minimum length* or *parsimony criterion*, provided that $\pi > \frac{1}{2}$. Various other parsimony models are derived or discussed in Camin and Sokal (1965; Camin-Sokal parsimony), Farris (1973), Le Quesne (1974; Dollo parsimony), Felsenstein (1983b, 1988), Sober (1985), Felsenstein and Sober (1986) and Goldman (1990). Practical optimization algorithms can be found in Fitch (1971), Hartigan (1973; Fitch's algorithm), Sankoff (1975), Day et al. (1986).

A combinatorial and probabilistic analysis of parsimony trees for randomized or uniformly distributed sequence data in the discrete space $\mathscr{A}^p$ is provided by Carter et al. (1990), Steel (1992) and Steel et al. (1992) who use the duality between labeled trees (with node labels from $\mathscr{A}^p$) and graph colouring problems (with $|\mathscr{A}|^p$ colours).

Sneath (1989) considers the random sampling of characters and shows that the probability of detecting the correct tree can be small if there are only few characters. Quite generally, the maximum likelihood and parsimony methods were questioned by Nei (1987) and Saitou (1988) under the aspect that the likelihood values are incomparable for different topologies T. There exist simple models where the results of a parsimony method will converge to a wrong phylogenetic tree for $n \to \infty$, even for equal mutation rates λ_l (Felsenstein, 1978; Saitou, 1988; Hendy and Penny, 1989). This has motivated the development of improved estimation methods for T based on 'invariant functions' of the distances which involve, e.g., the 'four-point inequality' which characterizes an additive tree (Lake, 1987; Cavender, 1991; Day, 1991; Navidi et al., 1993). In order to evaluate the confidence in estimated phylogenies Li (1989) proposes tests for the significance of the estimated internodal lengths of T, Kishino and Hasegawa (1989) consider the variance of the log likelihood ratio, and Hasegawa et al. (1988) and Felsenstein (1985) use bootstrap resampling for this purpose.

6.4. *Purely random hierarchies and trees*

While the previous subsections focussed on models for hierarchies originating from a 'natural'evolutionary clustering process there is also some need for the specification of models for 'purely random' hierarchies (or dendrograms): In fact, such concepts are indispensable if we want, e.g.,

(a) to decide if a calculated dendrogram points really to an underlying hierarchical structure of the data (as opposed to a situation where it bears no more structure than a 'purely random' dendrogram; e.g., Murtagh, 1983; Frank and Svensson, 1981), or

(b) to compare two dendrograms $(\mathscr{H}_1, h_1)$, $(\mathscr{H}_2, h_2)$ resp. δ_1, δ_2, two hierarchies $\mathscr{H}_1, \mathscr{H}_2$ or two phylogenetic trees and to check if they are more different from each other than to be expected under 'pure randomness' (eventually assuming the same clustering strategy; Lapointe and Legendre, 1992a, b), or, as a special case,

(c) investigate if a new dendrogram $(\mathscr{H}_1, h_1)$ obtained from the data is significantly different from a known (e.g., traditionally established) classification $(\mathscr{H}_2, h_2)$ of the objects (Nemec and Brinkhurst, 1988).

Similar questions relate to phylogenetic or additive trees as well as to 'non-weighted' modifications considering the tree topology only.

In situations like (b) or (c), the usual approach proceeds by defining, in a first step, a suitable consensus index or a distance measure $D(T_1, T_2)$ between two hierarchical structures T_1, T_2; typical examples are the *partition (or symmetric difference) metric*, *quartet metrics*, the *nearest-neighbour interchange (NNI) metric* and the *cophenetic (correlation) coefficient* (for details see, e.g., Boorman and Oliver, 1973; Waterman and Smith, 1978; Day 1983; Fowlkes and Mallows, 1983; Brown and Day, 1984; Lapointe and Legendre 1990, 1992b; Steel and Penny, 1993). Then, in a second step, we have to check if the observed value $D(T_1, T_2)$ for the calculated classifications T_1, T_2 is significantly larger than to be expected under a hypothesis H_0 of 'pure randomness' of T_1 and/or T_2 or of the data $x_1, \ldots, x_n$ or (d_{kl}) (see Shao and Rohlf, 1983; Shao and Sokal, 1986; Lapointe and Legendre, 1990). Such an approach requires the calculation of suitable percentage points or, at least, of the expectation and variance of the criterion D which proves to be quite difficult in the general case since the relation:

$$\textit{calculated} - \textit{classification} = \textit{clustering} - \textit{algorithm}(\text{data})$$

is usually too complex as to allow to obtain any exact null distribution for D. In most cases, randomization, permutation and Monte Carlo tests will be in order (Rohlf, 1965; Hubert, 1985; Archie, 1989) and percentage points of D are approximated by bootstrapping (Felsenstein, 1985; Sanderson, 1989; Krajewski and Dickerman, 1990) and simulation under H_0.

In this latter framework, there is a lot of essentially combinatorial approaches for defining 'purely random' hierarchies, dendrograms and trees by an equidistribution on a finite set $\mathscr{S}$ of 'distinguishable' hierarchical structures (hierarchies, dendrograms, trees, eventually including size and shape constraints). The investigation and simulation of these null models requires combinatorial enumeration techniques as well as computationally simple generation algorithms. This topic is discussed and surveyed, e.g., by Simberloff (1987), Furnas (1984), Quiroz (1989), Page (1991), Lapointe and Legendre (1991) and Steel and Penny (1993) who present many algorithms. In the following we will briefly sketch some typical cases.

We start with the remark that a dendrogram $(\mathscr{H}, h)$ for n objects can be considered as a rooted labeled and weighted tree T_n with n leaves and (at most) $n-1$ interior points (corresponding to $n-1$ cluster fusions). It induces (and is induced by) an additive (or path length) tree where each leaf has the same path distance from the root. Eliminating edge lengths and the root we obtain the combinatorial structure of an undirected tree-like graph. Obviously, two dendrograms or (additive) trees may differ in various aspects such as topology T, labeling of vertices L, fusion ranks R and

fusion levels (or edge weights) W and therefore the definition of a null distribution H_0 of 'pure randomness' requires a careful specification of the set $\mathscr{S}$ in order to include (only) the practically relevant aspects. In particular, we have to distinguish, between rooted and unrooted trees, unlabeled, fully and terminally labeled trees, binary and t-ary trees, ranked and weighted dendrograms (with rank-ordered resp. real-valued fusion levels) etc. which all define different levels of analysis. Exact enumerations and probabilities can be obtained in a few (unweigthed) cases, in particular for the size of $\mathscr{S}$:

(a) there are $a_n = n^{n-2}$ fully labeled unrooted trees with n vertices (Cayley, 1889);

(b) the number of rooted binary and terminally labeled (unweighted) trees (T, L) with n leaves and $n-1$ interior edges (characterizing the nesting structure in a bifurcating hierarchy $\mathscr{H}$) is given by $b_n = (2n-3)!/[2^{n-2}(n-2)!] = 1 \cdot 3 \cdot 5 \cdots (2n-3)$ (Harding, 1971);

(c) the number of unrooted binary and terminally labeled (unweighted) trees (unweighted phylogenetic trees) with n leaves is given by $c_n = b_n/(2n-3) = (2n-5) \cdot (2n-7) \cdots 5 \cdot 3 \cdot 1$ (Schröder, 1870);

(d) there are $d_n = n!(n-1)!/2^{n-1}$ topologically distinguishable binary ranked dendrograms (T, L, R) with n labeled leaves and $n-1$ distinct fusion ranks (Frank and Svensson, 1981).

For this latter random dendrogram case (d), Dale and Moon (1988) obtain the exact distribution of the size of the smaller subtree attached to the root, the number S of terminal single objects (i.e., the last S joins in the dendrogram join all a single object to non-singleton classes), and the number M_k of subtrees with $k+1$ leaves. Harding (1971) derives probabilities relating to the shape of rooted (unlabeled and labeled) binary trees, Day (1986) and Steel and Penny (1993) obtain the expectation, variance and (simulated or asymptotically Poisson) distributions for several tree comparison metrics under various specifications for H_0.

Since the number of trees or hierarchies is extensively large even for a small number n of objects, a full enumeration is not possible and therefore effective methods for the generation and simulation of random structures become very important: Furnas (1984) describes, e.g., a two-step procedure for generating random additive trees by first obtaining a random rooted binary and terminally labeled tree (T, L) (using an enumeration technique for the classical Prüfer code of a tree), and then assigning random lengths to its branches. Similarly, De Soete (1984) constructs a random binary dendrogram (T, L, W) by assigning random fusion level values to the vertices of (T, L) whilst Rohlf (1983) and Murtagh (1983) consider random binary *ranked* dendrograms (T, L, R) assuming $n-1$ different fusion ranks. Murtagh (1984) investigates a packed representation of the 'shape' (T, R) of labeled dendrogams (T, L, R). The method of Lapointe and Legendre (1991) starts with $n-1$ random uniformly distributed fusion level values $W_1, \ldots, W_{n-1} \in [0, 1]$, arranges them in an ultrametric distance matrix $(\delta_{kl})_{n \times n}$ and assigns random labels to its n rows (and columns) in order to construct a random dendrogram (with extensions for the case of other level distributions by using a double permutation method). The triple permutation algorithm in Lapointe and Legendre (1992a) constructs a random additive tree by means of the fact that any additive tree metric (a_{kl}) has a composition

$(a_{kl}) = (\delta_{kl}) + (\tau_{kl})$ with an ultrametric δ and a star metric τ. Finally, Quiroz (1989) describes the construction of random rooted fully or terminally labeled t-ary trees with several generalizations, and Van Cutsem (1993) proposes, in analogy to Harding (1971), an iterative Markovian bifurcating process for obtaining a random rooted and terminally labeled binary tree (which is implemented recursively by means of a suitable declarative computer language).

Lapointe and Legendre (1992a, b) have used these generation methods to derive simulated percentage points for various distance indices $D(T_1, T_2)$ (e.g., cophenetic correlation) and random dendrograms and/or additive trees T_1, T_2.

References

Anderson, J.J., Normal mixtures and the number of clusters problem, *Comput. Statist. Quart.*, **2** (1985) 3–14.

Archie, J.W., A randomization test for phylogenetic information in systematic data, *Systematic Zoology*, **38** (1989) 239–252.

Banks, D. and K. Carley, Metric inference for social networks, *J. Classification*, **11** (1994) 121–149.

Benkaraache, T. and B. Van Cutsem, Comparison of hierarchical classifications, Technical Report RT-100 (Laboratoire de Modélisation et Calcul, Institut IMAG, CNRS, Grenoble, 1993) 10 pp.

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

Bhattacharya, R.N. and J.K. Ghosh, A class of U-statistics and asymptotic normality of the number of k-clusters. *J. Multivariate Analysis*, **43** (1992) 300–330.

Bishop, M.J. and A.E. Friday, Evolutionary trees from nucleic acid and protein sequences, *Proc. Royal Soc. London*, B **226** (1985) 271–302.

Bock, H.H., The equivalence of two extremal problems and its application to the iterative classification of multivariate data, Report of the Conf. on "Medizinische Statistik" (Forschungsinstitut Oberwolfach, February 1969).

Bock, H.H., *Automatische Klassifikation* (*Clusteranalyse*) (Vandenhoeck and Ruprecht, Göttingen, 480 pp., 1974).

Bock, H.H., 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.

Bock, H.H., On some significance tests in cluster analysis, *J. Classification*, **2** (1985) 77–108.

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

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.

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

Bock, H.H., Probabilistic aspects in cluster analysis, in: O. Opitz (Ed.), *Conceptual and numerical analysis of data* (Springer, Heidelberg, 1989a) 12–44.

Bock, H.H., 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 1989b).

Bock, H.H., 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) Vol. II, 115–117.

Bock, H.H., Probabilistic approaches and hypotheses testing in partitioning cluster analysis, in: Ph. Arabie, G. De Soete and L. Hubert (Eds.), *Clustering and classification* (World Science Publishers, Singapore and River Edge, NJ/USA, 1996a) 377–453.

Bock, H.H., Probabilistic models in partitional cluster analysis, in: A. Ferligoj and A. Kramberger (Eds.), *Developments in data analysis* (Metodološki zvezki, 12, FDV, Ljubejana, 1996b).

Bock, H.H., Probability models for convex clusters, in: O. Opitz and R. Klar (Eds.), *Classification and data analysis* (Springer, Heidelberg, to appear), 11 pp. (1996c).

Bock, H.H. and P. Ihm (Eds.), *Classification, data analysis and knowledge organization* (Springer, Heidelberg, 1991).

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

Bock, H.H. and W. Polasek (Eds.), *Data analysis and information systems* (Springer, Heidelberg, 1996) 548 pp.

Bollobás, B., *Random graphs* (Academic Press, London, 1985).

Boorman, S.A. and D.C. Olivier, Metrics on spaces of finite trees, *J. Math. Psychology*, **10** (1973) 26–59.

Bozdogan, H., Choosing the number of clusters, subset selection of variables, and outlier detection in the standard mixture-model cluster analysis, in: E. Diday et al. (Eds.), 1994, 169–177.

Brown, E.K. and W.H.E. Day, A computationally efficient approximation to the nearest neighbor interchange metric, *J. Classification*, **1** (1984) 93–124.

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

Camin, J.H. and R.R. Sokal, A method for deducing branching sequences in phylogeny, *Evolution*, **19** (1965) 311–326.

Carter, M., M. Hendy, D. Penny, L.A. Székely and N.C. Wormald, On the distribution of lengths of evolutionary trees, *SIAM J. Discrete Math.*, **3** (1990) 38–47.

Cavender, J.A., Necessary conditions for the method of inferring phylogeny by linear invariants, *Molecular Biosciences*, **103** (1991) 69–75.

Cavender, J.A. and J. Felsenstein, Invariants of phylogenies in a simple case with discrete states, *J. Classification*, **4** (1987) 57–71.

Cayley, A., A theorem on trees, *Quarterly J. Pure Appl. Math.*, **23** (1889) 376–378.

Céleux, G. and G. Govaert, Clustering criteria for discrete data and latent class models, *J. Classification*, **8** (1991) 157–176.

Cressie, N., *Statistics for spatial data* (Wiley, New York, 1991).

Dale, M.R.T. and J.W. Moon, Statistical tests on two characteristics of the shapes of cluster diagrams, *J. Classification*, **5** (1988) 21–38.

Day, W.H.E., Distributions of distances between pairs of classifications, in: J. Felsenstein (Ed.), 1983c, 127–131.

Day, W.H.E., Analysis of quartet dissimilarity measures between undirected phylogentic trees, *Systematic Zoology*, **35** (1986) 325–333.

Day, W.H.E., Estimating phylogenies with invariant functions of data, in: H.H. Bock and P. Ihm (eds.), 1991, 248–253.

Day, W.H.E., D.S. Johnston and D. Sankoff, The computational complexity of inferring rooted phylogenies by parsimony, *Math. Biosciences*, **81** (1986) 33–42.

Degens, P.O., Hierarchical cluster methods as maximum likelihood estimators, in: J. Felsenstein (ed.), 1983c, 249–253.

Degens, P.O., Ultrametric approximation to distances, *Computat. Statist. Quart.*, **2** (1) (1985) 93–101.

Degens, P.O., Reconstruction of phylogenies by weighted genetic distances, in: H.H. Bock (Ed.), 1988, 727–739.

Degens, P.O., B. Lausen and W. Vach, Reconstruction of phylogenies by distance data: Mathematical framework and statistical analysis, Research report 88/8 (Department of Statistics, University of Dortmund, 1988) 34 pp.

De Soete, G., Ultrametric tree representations of incomplete similarity data, *J. Classification*, **1** (1984) 235–242.

De Soete, Tree representations of proximity data by least squares methods, in: H.H. Bock (Ed.), 1988, 147–157.

De Soete, J.D. Carroll and W.S. Desarbo, Least squares algorithms for constructing constrained ultrametric and additive tree representations of symmetric proximity data, *J. Classification*, **4** (1987) 155–173.

Dette, H. and N. Henze, The limit distribution of the largest nearest neighbour link in the unit d-cube, *J. Appl. Probab.*, **26** (1988).

Diday, E., Introduction à l'analyse factorielle typologique, (Rapport de Recherche No. 27, IRIA, LeChesnay, France, 1973).

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

Farris, J.S., A probability model for inferring phylogenetic trees, *Systematic Zoology*, **22** (1973) 250–256.

Felsenstein, J., Cases in which parsimony or compatibility methods will be positively misleading, *Systematic Zoology*, **27** (1978) 401–410.

Felsenstein, J., Evolutionary trees from DNA sequences: A maximum likelihood approach, *J. Molecular Evolution*, **17** (1981) 368–376.

Felsenstein, J., Statistical inference of phylogenies, *J. Royal Statist. Soc.*, A **146** (1983a) 246–272.

Felsenstein, J., Parsimony in systematics: Biological and statistical issues, *Ann. Rev. Ecological Systems*, **14** (1983b) 313–333.

Felsenstein, J., Ed., *Numerical taxonomy*, (NATO Advanced Studies Institute, Ser. G. (Ecological Sciences) 1. Springer, Berlin, 1983c).

Felsenstein, J., The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility, in: T. Duncan and T.F. Steussy (Eds.), *Cladistics: perspectives on the reconstruction of evolutionary history* (Columbia Univ. Press, New York, 1984) 169–191.

Felsenstein, J., Confidence limits on phylogenies: An approach using the bootstrap, *Evolution*, **39** (1985) 783–791.

Felsenstein, J., Phylogenies from molecular sequences: inference and reliability, *Annual Review of Genetics*, **22** (1988) 521–565.

Felsenstein, J., *PHYLIP* (*Phylogenetic inference package*), Version 3.3 (University of Washington, Seattle, 1990).

Felsenstein, J., and E. Sober, Parsimony and likelihood: an exchange, *Systematic Zoology*, **35** (1986) 617–626.

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

Fitch, W.M., Towards defining the course of evolution: Minimum change for a specific tree topology, *Systematic Zoology*, **20** (1971) 406–416.

Fowlkes, E.B. and C.L. Mallows, A method for comparing two hierarchical clusterings, *J. Amer. Statist. Assoc.*, **78** (1983) 553–569.

Frank, O., Inferences concerning cluster structure, in: COMPSTAT 1978 (Physica-Verlag, Würzburg, 1978) 259–265.

Frank, O. and D. Strauss, Markov graphs, *J. Amer. Statist. Assoc.*, **81** (1986) 832–842.

Frank, O. and K. Svensson, On probability distributions of single-linkage dendrograms, *J. Statist. Comput. Simul.*, **12** (1981) 121–131.

Furman, W.D. and B.G. Lindsay, Testing for the number of components in a mixture of normal distributions using moment estimators, *Comput. Statist. Data Anal.*, **17** (1994) 473–492.

Furnas, G.W., The generation of random, binary unordered trees, *J. Classification*, **1** (1984) 187–234.

Gaul, W. and D. Pfeifer (Eds.), *From data to knowledge. Proc. 18th Ann. Conf. Gesellschaft für Klassifikation*, Oldenburg, 1994 (Springer-Verlag, Heidelberg, 1996) 472 pp.

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

Godehardt, E., Multigraphs for the uncovering and testing of structures, in: H.H. Bock and P. Ihm (Eds.), 1991, 43–52.

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

Goldman, N., Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses, *Systematic Zoology*, **39** (1990) 345–361.

Gordon, A.D., Clustering algorithms and cluster validation, in: P. Dirschedl and R. Ostermann (Eds.), *Computational Statistics* (Physica-Verlag, Heidelberg, 1994), 497–512.

Gordon, A.D., Null models in cluster validation, in: W. Gaul and D. Pfeifer (Eds.), 1996a, 32–44.

Gordon, A.D., Hierarchical classification, in: P. Arabie, L.J. Hubert and G. De Soete (Eds.), *Clustering and classification* (World Science Publishers, Singapore and River Edge, NJ/USA, 1996b) 65–121.

Guénoche, A., Hiérarchies conceptuelles de données binaires, *Math. Inform. Sci. Hum.*, **121** (1993a) 23–34.

Guénoche, A., Alignment and hierarchical clustering method for strings, in: O. Opitz, B. Lausen and R. Klar (Eds.), *Information and classification: concepts, methods and applications* (Springer, Heidelberg, 1993b) 403–412.

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

Hansen, P. and B. Jaumard, Computational methods in clustering from a mathematical programming viewpoint, in: H.H. Bock and W. Polasek (Eds.), 1996, 24–40.

Harding, E.F., The probabilities of rooted tree-shapes generated by random bifurcation, *Advances in Applied Probability*, **3** (1971) 44–77.

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

Hartigan, J.A., Mimimum mutation fits to a given tree, *Biometrics*, **29** (1973) 53–65.

Hartigan, J.A., Asymptotic distributions for clustering criteria, *Ann. Statist.*, **6** (1978) 117–131.

Hartigan, J.A., The SPAN test for multimodality, in: H.H. Bock (Ed.), 1988, 229–236.

Hartigan, J.A. and P.M. Hartigan, The DIP test for multimodality, *Ann. Statist.*, **13** (1985) 70–84.

Hartigan, J.A. and S. Mohanty, The RUNT test for multimodality, *J. Classification*, **9** (1992) 63–70.

Hasegawa, M., H. Kishino and T. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *J. Molecular Evolution*, **22** (1985) 160–174.

Hasegawa, M., H. Kishino and T. Yano, Phylogenetic inference from DNA sequence data, in: K. Matusita (Ed.), *Statistical theory and data analysis II* (North Holland, Amsterdam, 1988) 1–13.

Hendy, M.D. and D. Penny, Branch and bound algorithms to determine evolutionary trees, *Math. Biosciences*, **59** (1982) 277–290.

Hendy, M.D. and D. Penny, A framework for the quantitative study of evolutionary trees, *Systematic Zoology*, **38** (1989) 297–309.

Henze, N., The limit distribution for maxima of "weighted" rth-nearest neighbour distances, *J. Appl. Probab.*, **19** (1982) 344–354.

Hoffman, R. and A.K. Jain, A test of randomness based on the minimal spanning tree, *Pattern Recognition Letters*, **1** (1983) 175–180.

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

Hubert, L., *Assignment methods in combinatorial data analysis* (Marcel Dekker, New York, 1987).

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

Jank, W., Zur Untersuchung des Varianzkriteriums in der Clusteranalyse: Optimale und Stationäre Partitionen des $\mathbb{R}^2$ und die Verteilung zugehöriger Gütekriterien (Diploma thesis, Institute of Statistics, Technical University of Aachen, Germany) 204 pp.

Janson, S., Maximal spacings in several dimensions, *Ann. Probab.*, **15** (1987) 274–280.

Kimura, M., A simple method for estimating evolutionary rates of base substitutions trough comparative studies of nucleotide sequences, *J. Molecular Evolution*, **2** (1980) 87–90.

Kishino, H. and M. Hasegawa, Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of Hominoidea, *J. Molecular Evolution*, **29** (1989) 170–179.

Krajowski, C. and A.W. Dickerman, Bootstrap analysis of phylogenetic trees derived from DNA hybridization distances, *Systematic Zoology*, **39** (1990) 383–390.

Lake, J.A., A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony, *Molecular Biology and Evolution*, **4** (1987) 167–191.

Lapointe, F.-J. and P. Legendre, A statistical framework to test the consensus of two nested classifications, *Systematic Zoology*, **39** (1990) 1–13.

Lapointe, F.-J. and P. Legendre, The generation of random ultrametric matrices representing dendrograms, *J. Classification*, **8** (1991) 177–200.

Lapointe, F.-J. and P. Legendre, A statistical framework to test the consensus among additive trees (cladograms), *Systematic Biology*, **41** (1992a) 158–171.

Lapointe, F.-J. and P. Legendre, Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees, *Systematic Biology*, **41** (1992b) 378–384.

Lausen, B., Exploring homologuous tRNA sequence data: Positional mutation rates and genetic distance, in: O. Opitz (Ed.), *Conceptual and numerical analysis of data* (Springer, Heidelberg, 1989) 481–488.

Lausen, B., Statistical analysis of genetic distance data, in: H.H. Bock and P. Ihm (Eds.), 1991, 254–261.

Lausen, B. and P.O. Degens, Variance estimation and the reconstruction of phylogenies, in: P.O. Degens, H.-J. Hermes and O. Opitz (Eds.), *Classification and its environment. Studien zur Klassifikation SK–17* (Indeks-Verlag, Frankfurt a. M., 1986) 306–314.

Lausen, B. and P.O. Degens, Evaluation of the reconstruction of phylogenies with DNA–DNA hybridization, in: H.H. Bock (Ed.), 1988, 367–374.

Le Quesne, W.J., The uniquely involved character concept and its cladistic application, *Systematic Zoology*, **23** (1974) 513–517.

Lee, K. L., Multivariate tests for clusters, *J. Amer. Statist. Assoc.*, **74** (1979) 708–714.

Li, W.-H., A statistical test of phylogenies estimated from sequence data, *Molecular Biology and Evolution*, **6** (1989) 424–435.

Ling, R.F., A probability theory of clustering, *J. Amer. Statist. Assoc.*, **68** (1973) 159–164.

Ling, R.F., An exact probability distribution on the connectivity of random graphs, *J. Math. Psychology*, **12** (1975) 90–98.

Ling, R. F. and G.G. Killough, Probability tables for cluster analysis based on a theory of random graphs, *J. Amer. Statist. Assoc.*, **71** (1976) 293–300.

Matula, D.W., On the complete subgraphs of a random graph, *Combinatorial mathematics and its applications* (Chapel Hill, N.C., 1970) 356–369.

Matula, D.W., The employee party problem, *Notices Amer. Math. Soc.*, **19** (1972) A–382.

Matula, D.W., The largest clique in a random graph, Technical report CS–7608 (Dept. Computer Science, Southern Methodist Univ., Dallas TX, April 1976) 22 pp.

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

Mountford, M.D., A test of the difference between clusters, in: G.P. Patil et al. (Eds.), *Statistical ecology. Vol. 3* (Pennsylvania State Univ. Press, University Park, Pa., 1970) 237–257.

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

Murtagh, F., A probability theory of hierarchic clustering using random dendrograms, *J. Statist. Comput. Simul.*, **18** (1983) 145–157.

Murtagh, F., Counting dendrograms: A survey, *Discrete Appl. Math.*, **7** (1984) 191–199.

Navidi, W.C., G.A. Churchill and A. von Haeseler, Phylogenetic inference: Linear invariants and maximum likelihood, *Biometrics*, **49** (1993) 543–555.

Nemec, A.F.L. and R.O. Brinkhurst, The Fowlkes-Mallows statistic and the comparison of two independently determined dendrograms, *Canad. J. Fish. Aquat. Sci.*, **45** (1988) 971–975.

Nowicki, K., Asymptotic Poisson distributions with applications to statistical analysis of graphs, *Advances Appl. Probab.*, **20** (1988) 315–330.

Page, R.D.M., Random dendrograms and null hypotheses in cladistic biogeography, *Systematic Zoology*, **40** (1991) 54–62.

Palmer, E.M., *Graphical evolution: An introduction to the theory of random graphs* (Wiley, New York, 1985).

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

Pociecha, J. and A. Sokolowski, Empirical tests of multidimensional uniformity, *Control and Cybernetics*, **18** (1) (1989) 81–86.

Pollard, D., A central limit theorem for k-means clustering, *Ann. Probab.*, **10** (1982) 919–926.

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

Quiroz, A.J., Fast random generation of binary, t-ary and other types of trees, *J. Classification*, **6** (1989) 223–231.

Rasson, J.P., A. Hardy and D. Weverbergh, Point process, classification and data analysis, in: H.H. Bock (Ed.), 1988, 245–256.

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

Ripley, B.D., *Spatial statistics* (Wiley, New York, 1981).

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

Rohlf, F.J., A randomization test of the nonspecificity hypothesis in numerical taxonomy, *Taxon*, **14** (1965) 262–267.

Rohlf, F.J., Numbering binary trees with labeled terminal vertices, *Bull. Math. Biology*, **45** (1983) 33–40.

Rozál, G.P.M. and J.A. Hartigan, The MAP test for multimodality, *J. Classification*, **11** (1994) 3–36.

Saitou, N., Property and efficiency of the maximum likelihood method for molecular phylogeny, *J. Molecular Evolution*, **27** (1988) 261–273.

Sanderson, M.J., Confidence limits on phylogenies: The Bootstrap revisited, *Cadistics*, **5** (1989) 113–129.

Sankoff, D., Minimal mutation trees of sequences, *SIAM J. Appl. Math.*, **28** (1975) 35–42.

Sawitzki, G., The excess-mass approach and the analysis of multi-modality, in: W. Gaul and D. Pfeifer (Eds.), 1996, 203–211.

Sbihi, A. and J.-G. Postaire, Mode extraction by multivalue morphology for cluster analysis, in: W. Gaul and D. Pfeifer (Eds.), 1996, 212–221.

Schöniger, M., A. Janke and A. von Haeseler, How to deal with third codon positions in phylogenetic analysis, in: H.H. Bock, W. Lenski, M.M. Richter (Eds.), 1994, 376–383.

Schröder, E., Vier combinatorische Probleme, *Zeitschrift für Mathematik und Physik*, **15** (1870) 361–376.

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

Shao, K. and F.J. Rohlf, Sampling distribution of consensus indices when all bifurcating trees are equally likely, in: J. Felsenstein (Ed.), 1983c, 132–137.

Shao, K. and R.R. Sokal, Significance tests of consensus indices, *Systematic Zoology*, **35** (1986) 582–590.

Silverman, B.W., Using kernel density estimates to investigate multimodality, *J. Royal Statist. Soc.*, B **43** (1981) 97–99.

Silverman, B.W. and T. Brown, Short distances, flat triangles and Poisson limits, *J. Appl. Probab.*, **15** (1978) 816–826.

Sneath, P.H.A., Predictivity in taxonomy and the probability of a tree, *Pl. Syst. Evol.*, **167** (1989) 43–57.

Simberloff, D., Calculating the probabilities that cladograms match: A method of biogeographic inference, *Systematic Zoology*, **36** (1987) 175–195.

Sober, E., A likelihood justification of parsimony, *Cladistics*, **1** (1985) 209–233.

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

Sriram, N. and S. Lewis, Constructing optimal ultrametrics, *J. Classification*, **10** (1993) 241–268.

Steel, M.A., Distributions on bicoloured binary trees from the principle of parsimony, *SIAM J. Discrete App. Math.*, 1992.

Steel, M. and D. Penny, Probability distributions of tree comparison metrics-some new results, *Systematic Biology*, **42** (1993).

Steel, M.A., M.D. Hendy and D. Penny, Significance of the length of the shortest tree, *J. Classification*, **9** (1992) 71–90.

Steele, J.M., Growth rates of euclidean minimal spanning trees with power weighted edges, *Annals of Probability*, **16** (1988) 1767–1787.

Steele, J.M. and L. Tierney, Boundary domination and the distribution of the largest nearest neigbor link in higher dimensions, *J. Appl. Probab.*, **23** (1988) 524–528.

Tabakis, E., On the longest edge of the minimal spanning tree, in: W. Gaul and D. Pfeifer (Eds.), 1996, 222–230.

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

Van Cutsem, B., Combinatorial structures and structures for classification, *Comput. Statistics and Data Analysis*, **23** (1996) This volume.

Van Cutsem, B. and B. Yeart, Probability distributions on indexed dendrograms and related problems of classifiability, in: H.H. Bock and W. Polasek (Eds.), 1996, 73–87.

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

Waterman, M.S. and T.F. Smith, On the similarity of dendrograms, *J. Theoret. Biology*, **73** (1978) 789–800.

Windham, M.P., Parameter modification for clustering criteria, *J. Classification*, **4** (1987) 191–214.

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

Windham, M.P. and A. Cutler, Mixture analysis with noisy data, in: E. Diday et al. (Eds.), 1994, 155–160.

Wolf, K. and P.O. Degens, Variance estimation in the additive tree model, in: H.H. Bock and P. Ihm (Eds.), 1991, 262–269.