J. R. Statist. Soc. B (2004)
66, *Part* 4, *pp.* 815–849

Clustering objects on subsets of attributes

Jerome H. Friedman

Stanford University, USA

and Jacqueline J. Meulman

Leiden University, the Netherlands

[*Read before* The Royal Statistical Society *at a meeting organized by the Research Section on Wednesday, May 5th, 2004*, Professor J. T. Kent *in the Chair*]

Summary. A new procedure is proposed for clustering attribute value data. When used in conjunction with conventional distance-based clustering algorithms this procedure encourages those algorithms to detect automatically subgroups of objects that preferentially cluster on *subsets* of the attribute variables rather than on all of them simultaneously. The relevant attribute subsets for each individual cluster can be different and partially (or completely) overlap with those of other clusters. Enhancements for increasing sensitivity for detecting especially low cardinality groups clustering on a small subset of variables are discussed. Applications in different domains, including gene expression arrays, are presented.

Keywords: Bioinformatics; Clustering on variable subsets; Distance-based clustering; Feature selection; Gene expression microarray data; Genomics; Inverse exponential distance; Mixtures of numeric and categorical variables; Targeted clustering

1. Introduction

The goal of cluster analysis is to partition a data set of N objects into subgroups such that those in each particular group are more similar to each other than to those of other groups. Defining an 'encoder' function $c(i)$ that maps each object i to a particular group G_l $(1 \leqslant l \leqslant L)$

$$c(i) = l \Rightarrow i \in G_l, \tag{1}$$

we can formalize this goal as finding the 'optimal' encoder $c^*(i)$ that minimizes a criterion $Q(c)$ that measures the degree to which the goal is not being met:

$$c^* = \arg \min_c \{Q(c)\}. \tag{2}$$

One such criterion is

$$Q(c) = \sum_{l=1}^{L} \frac{W_l}{N_l^2} \sum_{c(i)=l} \sum_{c(j)=l} D_{ij}, \tag{3}$$

where D_{ij} is a defined distance or dissimilarity measure between every pair of objects (i, j) and N_l is the number of objects assigned to the lth group,

$$N_l = \sum_{i=1}^{N} I\{c(i) = l\}, \tag{4}$$

Address for correspondence: Jerome H. Friedman, Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94305, USA.
E-mail: jhf@stanford.edu

1369–7412/04/66815

where the 'indicator' function $I(\cdot) \in \{0, 1\}$ indicates truth of its argument, and where $\{W_l\}_1^L$ in equation (3) are cluster weights. Thus criterion (3) is a weighted average over the groups, of the within-group mean distance between pairs of objects assigned to the same group. The cluster weights $\{W_l\}_1^L$ are taken to be functions of the group sizes $\{N_l\}_1^L$ and can be used to regulate the distribution of groups sizes of the solution (2). (See Hubert *et al.* (2001), page 19, for a review of possible heterogeneity measures within a subset.) The usual choice $\{W_l = N_l^2\}_1^L$ gives the same influence to all object pairs in the criterion (3), encouraging equal-sized solution clusters.

2. Attribute value data

When each object i is characterized by a set of n measured attributes (variables),

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ik}, \ldots, x_{in}),$$

distances between pairs of objects D_{ij} as in equation (3) are based on their respective values (x_{ik}, x_{jk}) on each attribute k. A well-known example is the Gower (1971) coefficient of similarity. One defines a distance d_{ijk} between objects (i, j) separately on each attribute k, and then D_{ij} is taken to be a (weighted) average of the respective attribute distances

$$D_{ij} = \sum_{k=1}^{n} w_k d_{ijk} \tag{5}$$

with

$$\{w_k \geqslant 0\}_1^n \text{ and } \sum_{k=1}^{n} w_k = 1. \tag{6}$$

For example, the individual attribute distances can be taken as

$$d_{ijk} = \delta_{ijk}/s_k \tag{7}$$

where for numeric-valued attributes

$$\delta_{ijk} = |x_{ik} - x_{jk}|, \tag{8}$$

or often its square, and for categorically valued (nominal) attributes

$$\delta_{ijk} = I(x_{ik} \neq x_{jk}). \tag{9}$$

There are numerous suggestions in the literature for distance measures on individual attributes other than equations (8) and (9). Particular choices reflect the goal of the cluster analysis. The approach that is presented in this paper applies to any such definitions. The denominator s_k in equation (7) provides a scale for measuring 'closeness' on each attribute. It is often taken to be

$$s_k = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ijk} \tag{10}$$

or some other measure of spread or dispersion of the $\{x_{ik}\}_{i=1}^N$ values over all objects. For equal attribute weights $\{w_k = 1/n\}_1^n$, this gives the same influence to all the attributes in defining criterion (3) and thereby on the solution (2). Sometimes the weights in definition (5) are set to unequal values to refine relative influence based on user domain knowledge or intuition further, if it is suspected that particular attributes are more relevant than others to clustering the objects.

From equations (3) and (5) we can express the (equal weight) clustering criterion as

$$Q(c)=\sum_{l=1}^{L} W_l\left(\frac{1}{n}\sum_{k=1}^{n} S_{kl}\right) \tag{11}$$

where

$$S_{kl}=\frac{1}{N_l^2}\sum_{c(i)=l}\sum_{c(j)=l} d_{ijk} \tag{12}$$

is a measure of the dispersion (scale) of the data values on the kth attribute for objects in the lth group, $\{x_{ik}|c(i)=l\}$. For example, if we use $d_{ijk}=(x_{ik}-x_{jk})^2/s_k^2$ then $S_{kl}=2\,\mathrm{var}\{x_{ik}/s_k|c(i)=l\}$. Thus, using equation (5) to define distance encourages expression (2)–(3) to seek clusters of objects that simultaneously have small dispersion on all or at least many of the attributes, i.e the objects within each solution subgroup are simultaneously close on a large number of the attributes.

3. Feature selection

Defining clusters in terms of simultaneous closeness on all attributes may sometimes be desirable, but often it is not. In data mining applications, the values of many attributes are often measured and it is unlikely that natural groupings will exist based on a large number of them. Usually, clustering, if it exists, occurs only within a relatively small unknown subset of the attributes. To the extent that all the attributes have equal influence, this type of clustering will be obscured and difficult to uncover.

The relative influence of each attribute x_k is regulated by its corresponding weight w_k in equation (5). Formally, feature selection seeks to find an optimal weighting $\mathbf{w}=\{w_k\}_1^n$ as part of the clustering problem by jointly minimizing the clustering criterion according to equations (3) and (5) with respect to the encoder c and weights $\mathbf{w}$, i.e.

$$(c^*,\mathbf{w}^*)=\arg\min_{(c,w)}\{Q(c,\mathbf{w})\} \tag{13}$$

where

$$Q(c,\mathbf{w})=\sum_{l=1}^{L}\frac{W_l}{N_l^2}\sum_{c(i)=l}\sum_{c(j)=l} D_{ij}[\mathbf{w}], \tag{14}$$

and $D_{ij}[\mathbf{w}]$ is given by equation (5), here emphasizing its dependence on the weights. The solution $\mathbf{w}^*$ has high weight values for those attributes that most exhibit clustering on the objects, and small values for those that do not participate in the clustering. The solution encoder c^* identifies the corresponding clusters. There is a vast literature on feature weighting and selection in clustering and classification; among others, see DeSarbo *et al.* (1984), De Soete *et al.* (1985), De Soete (1986, 1988), Fowlkes *et al.* (1988), Milligan (1989), Van Buuren and Heiser (1989), Gnanadesikan *et al.* (1995) and Brusco and Cradit (2001).

4. Clustering on different subsets of attributes

Although feature selection is often helpful, it only seeks groups that all cluster on the same subset of attributes. Those are attributes with large solution weight values (13). However, individual clusters may represent groupings on different (possibly overlapping) attribute subsets, and it is of interest to discover such structure. With feature selection, clustering on *different* subsets of attributes will still be obscured and difficult to uncover.

We can generalize criterion (14) to find clusters on separate attribute subsets by defining a separate attribute weighting $\mathbf{w}_l=\{w_{kl}\}_{k=1}^n$ for each individual group G_l, and jointly minimizing with respect to the encoder and all the separate weight sets associated with the respective groups, i.e.

$$(c^*,\{\mathbf{w}_l^*\}_1^L)=\arg\min_{(c,\{\mathbf{w}_l\}_1^L)}[Q(c,\{\mathbf{w}_l\}_1^L)], \tag{15}$$

where

$$Q(c,\{\mathbf{w}_l\}_1^L)=\sum_{l=1}^L \frac{W_l}{N_l^2}\sum_{c(i)=l}\sum_{c(j)=l} D_{ij}[\mathbf{w}_l], \tag{16}$$

and

$$D_{ij}[\mathbf{w}_l]=\sum_{k=1}^n w_{kl}d_{ijk}. \tag{17}$$

As before, expression (6), the attribute weights satisfy

$$\{w_{kl}\geqslant 0\}_1^n \text{ and } \sum_{k=1}^n w_{kl}=1, \qquad 1\leqslant l\leqslant L. \tag{18}$$

For any given encoder c, the solution to expression (15)–(16) for the corresponding attribute weights is $w_{kl}^*=I(k=k_l^*)$ where $k_l^*=\arg\min_{1\leqslant k\leqslant n}(S_{kl})$, with S_{kl} given by equation (12), i.e. the solution will put maximal (unit) weight on that attribute with smallest dispersion within each group G_l, and zero weight on all other attributes regardless of their respective dispersions within the group. Therefore, minimizing criterion (16) will produce solution groups that tend to cluster only on a *single* attribute. This type of clustering can be detected by simple inspection of the marginal data distributions on each attribute separately. Our goal is finding groups of objects that simultaneously cluster on subsets of attributes, where each subset contains more than one attribute.

This goal can be accomplished by modifying criterion (16) with an incentive (negative penalty) for solutions involving more attributes. One such incentive is the negative entropy of the weight distribution for each group

$$e(\mathbf{w}_l)=\sum_{k=1}^n w_{kl}\log(w_{kl}). \tag{19}$$

This function achieves its minimum value for equal weights and is correspondingly larger as the weights become more unequal. Incorporating equation (19), the modified criterion becomes

$$Q(c,\{\mathbf{w}_l\}_1^L)=\sum_{l=1}^L \frac{W_l}{N_l^2}\sum_{c(i)=l}\sum_{c(j)=l} D_{ij}^{(\lambda)}[\mathbf{w}_l], \tag{20}$$

with

$$D_{ij}^{(\lambda)}[\mathbf{w}_l]=\sum_{k=1}^n\{w_{kl}d_{ijk}+\lambda w_{kl}\ \log(w_{kl})\}+\lambda\ \log(n). \tag{21}$$

(The last term simply provides a translation so that $\min_{\mathbf{w}_l}(D_{ij}^{(\lambda)}[\mathbf{w}_l])=0$ whenever $\{d_{ijk}=0\}_{k=1}^n$.) The quantity $\lambda\geqslant 0$ controls the strength of the incentive for clustering on more attributes. It is a meta-parameter of the procedure and provides control over the type of clustering that is sought. Increasing or decreasing its value will encourage clusters on more or less attributes.

For a given encoder c, the solution to equation (15), minimizing expression (20)–(21) for the corresponding optimizing weight values is

$$w_{kl} = \exp\left(-\frac{S_{kl}}{\lambda}\right) \Big/ \sum_{k'=1}^{n} \exp\left(-\frac{S_{k'l}}{\lambda}\right), \tag{22}$$

with S_{kl} given by equation (12). This solution puts increased weight on attributes with smaller dispersion within each group G_l, where the degree of this increase is controlled by the value of λ. Setting $\lambda = 0$ places all weight on the attribute k with smallest S_{kl}, whereas $\lambda = \infty$ forces all attributes to be given equal weight for each group G_l.

Since all individual attribute distances (7) are normalized as in equation (10), the quantity S_{kl} will tend to be near 1 for attributes that do not contribute to the clustering of group G_l, and smaller for those that do. In this sense the value that is chosen for λ defines the meaning of 'clustering' on an attribute. A group of objects $G_l = \{i \,|\, c(i) = l\}$ is said to 'cluster' on attribute k, if S_{kl} for group G_l is smaller than the value of λ. Considerations governing the choice for its value are discussed in Section 6.1 later.

For a given encoder c, we can minimize expression (20)–(21) with respect to all the weights $\{\mathbf{w}_l\}_1^L$ (22), thereby producing a criterion $Q(c)$ that depends only on the encoder. The result is

$$Q(c) = \sum_{l=1}^{L} W_l \left[-\lambda \log \left\{ \frac{1}{n} \sum_{k=1}^{n} \exp\left(-\frac{S_{kl}}{\lambda}\right) \right\} \right], \tag{23}$$

where the optimal encoder is given by equation (2). The bracketed quantity in equation (23) is proportional to a generalized (Orlicz) mean

$$f^{-1}\left\{ \frac{1}{n} \sum_{k=1}^{n} f(S_{kl}) \right\} \tag{24}$$

of $\{S_{kl}\}_{k=1}^{n}$, where here

$$f(z) = \frac{1}{\exp(z/\lambda)} \tag{25}$$

is the inverse exponential function with scale parameter λ. This criterion (23) can be contrasted with that for ordinary clustering (11). Clustering based on distances using definition (5) with equal (or other prespecified) attribute weights minimizes the *arithmetic* mean of the attribute dispersions within each cluster; separate optimal attribute weighting within each cluster of objects minimizes the *inverse exponential* mean (24)–(25).

5. Search strategy

Defining the clustering solution as the minimum of some criterion does not fully solve the problem. We need a method for finding the minimizing encoder c^* that identifies the solution clusters. This is a combinatorial optimization problem (among others, see Hansen and Jaumard (1997), Hubert *et al.* (2001) and Van Os (2001)) for which a complete enumeration search over all possible encoders is computationally impractical for large problems. For these we must employ less than thorough heuristic search strategies.

For ordinary clustering based on criterion (3) or similar criteria, a large number of heuristic search strategies have been proposed. These are known as distance-based 'clustering algorithms' (for example, see Hartigan (1975), Späth (1980), Jain and Dubes (1988), Kaufman and Rousseeuw (1990), Arabie *et al.* (1996), Mirkin (1996) and Gordon (1999)). For attribute value data

(Section 2), clustering algorithms equivalently attempt to minimize criterion (11) by using equation (5) with equal (or prespecified) weights to define the distances D_{ij} between object pairs.

Criterion (23) is a more complicated highly non-convex function of the $\{S_{kl}\}$. The approach that is used here is to apply an alternating optimization strategy based on criterion (20). We start with an initial guess for the weight values, e.g. all values equal $\{w_{kl}=1/n\}$. Criterion (20) is then minimized with respect to the encoder given those weight values. Given that encoder, criterion (20) is minimized with respect to the weights, producing a new set of values for $\{\mathbf{w}_l\}_1^L$. These are then used to solve for a new encoder, and so on. This iterative procedure is continued until a (local) minimum is reached.

From equation (21) criterion (20) can be expressed as

$$Q(c,\{\mathbf{w}_l\}_1^L)=\sum_{l=1}^{L}\frac{W_l}{N_l^2}\sum_{c(i)=l}\sum_{c(j)=l}D_{ij}[\mathbf{w}_l]+\lambda\sum_{l=1}^{L}W_l\sum_{k=1}^{n}w_{kl}\,\log(w_{kl}), \tag{26}$$

where $D_{ij}[\mathbf{w}_l]$ is given by equation (17).

For a given encoder $c(\cdot)$, the minimizing solution for the weights is given by equation (22). Given a set of weight values $\mathbf{W}=\{\mathbf{w}_l\}_1^L\in R^{n\times L}$, the solution encoder $c^*(\cdot|\mathbf{W})$ minimizes

$$Q(c|\mathbf{W})=\sum_{l=1}^{L}\frac{W_l}{N_l^2}\sum_{c(i|\mathbf{W})=l}\sum_{c(j|\mathbf{W})=l}D_{ij}[\mathbf{w}_l]. \tag{27}$$

The form of this criterion is similar to that of criterion (3) where distance between objects assigned to the same group, $c(i|\mathbf{W})=c(j|\mathbf{W})=l$, is given by $D_{ij}[\mathbf{w}_l]$. However, conventional clustering algorithms cannot be directly used to attempt to minimize criterion (27) since they require distances to be defined between all object pairs, not just those assigned to the same group. The strategy that is employed here is to define a distance $D_{ij}[\mathbf{W}]$ between *all* object pairs that when used with standard clustering algorithms produces an encoder that approximates the solution $c^*(\cdot|\mathbf{W})$ minimizing criterion (27).

The starting-point for deriving such a distance measure is the assumption that $c^*(\cdot|\mathbf{W})$ has the property

$$\frac{1}{N_l^2}\sum_{c^*(i|\mathbf{W})=l}\sum_{c^*(j|\mathbf{W})=l}D_{ij}[\mathbf{w}_l]<\frac{1}{N_lN_m}\sum_{c^*(i|\mathbf{W})=l}\sum_{c^*(j|\mathbf{W})=m}D_{ij}[\mathbf{w}_l], \qquad m\neq l, \tag{28}$$

for all solution groups $G_l=\{i|c^*(i|\mathbf{W})=l\}$, i.e. the average distance between pairs of objects *within* the same group G_l, based on the weights for that group $\mathbf{w}_l$, is smaller than the corresponding average distance *between* groups based on $\mathbf{w}_l$. If this were not so, the value of criterion (27) could be further reduced by merging G_l with all groups G_m $(m\neq l)$ for which inequality (28) was violated. Furthermore from inequality (28) we have

$$\frac{1}{N_l^2}\sum_{c^*(i|\mathbf{W})=l}\sum_{c^*(j|\mathbf{W})=l}D_{ij}[\mathbf{w}_l]<\frac{1}{N_lN_m}\sum_{c^*(i|\mathbf{W})=l}\sum_{c^*(j|\mathbf{W})=m}\max(D_{ij}[\mathbf{w}_l],D_{ij}[\mathbf{w}_m]) \tag{29}$$

for $m\neq l$. Therefore, defining

$$D_{ij}^{(1)}[\mathbf{W}]=\max(D_{ij}[\mathbf{w}_{c(i|\mathbf{W})}],D_{ij}[\mathbf{w}_{c(j|\mathbf{W})}]) \tag{30}$$

we have

$$\frac{1}{N_l^2}\sum_{c^*(i|\mathbf{W})=l}\sum_{c^*(j|\mathbf{W})=l}D_{ij}[\mathbf{w}_l]=\frac{1}{N_l^2}\sum_{c^*(i|\mathbf{W})=l}\sum_{c^*(j|\mathbf{W})=l}D_{ij}^{(1)}[\mathbf{W}], \tag{31}$$

and from inequality (29)

$$\frac{1}{N_l^2} \sum_{c^*(i|\mathbf{W})=l} \sum_{c^*(j|\mathbf{W})=l} D_{ij}^{(1)}[\mathbf{W}] < \frac{1}{N_l N_m} \sum_{c^*(i|\mathbf{W})=l} \sum_{c^*(j|\mathbf{W})=m} D_{ij}^{(1)}[\mathbf{W}], \qquad m \neq l, \tag{32}$$

i.e. the solution encoder $c^*(\cdot|\mathbf{W})$ minimizing criterion (27) has the property that the average within-group distance, using equation (30), is smaller than the corresponding between-group average. The solutions that are produced by standard clustering algorithms also attempt to achieve this goal. Therefore, applying a standard clustering algorithm based on $D_{ij}^{(1)}[\mathbf{W}]$ (30) will attempt to produce a solution minimizing criterion (27).

The distance that is defined by equation (30) is not the only one that satisfies properties (31) and (32). Any 'majorizing' distance that is equal to $D_{ij}^{(1)}[\mathbf{W}]$ when $c(i|\mathbf{W}) = c(j|\mathbf{W})$ and is larger otherwise will share these properties. An example is

$$D_{ij}^{(2)}[\mathbf{W}] = \sum_{k=1}^{n} \max(w_{k,c(i|\mathbf{W})}, w_{k,c(j|\mathbf{W})}) d_{ijk}. \tag{33}$$

Any such distance could be used to produce a surrogate criterion for criterion (27) in the form of criterion (3) to be minimized by conventional clustering algorithms. A specific choice will depend on performance in the context of a particular clustering algorithm. This situation is common in optimization problems using heuristic search strategies, where we often choose to optimize a surrogate criterion with the same solution as the desired criterion. The choice of a surrogate is based solely on performance in the context of the search strategy chosen. Empirical evidence so far suggests that both definitions (30) and (33) yield similar results using common clustering algorithms, with definition (33) sometimes providing superior performance.

In summary, an alternating optimization algorithm attempting to minimize criterion (23) would initialize all weight values to $\mathbf{W} = \{w_{kl} = 1/n\}$. A solution encoder $c^*(\cdot|\mathbf{W})$ is obtained by applying a clustering algorithm using either equation (30) or equation (33) to define interpoint distances. New weight values $\mathbf{W}$ are computed based on $c^*(\cdot|\mathbf{W})$ using equations (12) and (22). These weight values define new interobject distances for the clustering algorithm. These steps are iterated until the solution stabilizes.

6. Weighted inverse exponential distance

The alternating optimization strategy that was outlined in the previous section is unlikely to produce satisfactory results if applied straightforwardly. The highly non-convex nature of criterion (23) induces a very large number of distinctly suboptimal local solutions. If the initial weight values $\mathbf{W} = \{1/n\}$ are far from their (global) minimizing values, it is likely that the alternating strategy will converge to one of these suboptimal local solutions. This will especially be so when there is clustering on small subsets of the attributes. To be successful, it is necessary either to find good initial weight values close to the solution values, or to use an alternative surrogate criterion for which the weight values $\mathbf{W} = \{1/n\}$ provide a good starting-point. Since it is usually difficult to assign good starting values without knowing the ultimate solution, the latter strategy is pursued here.

For any set of weights $\mathbf{w} = \{w_k\}_1^n$ consider the interpoint distance measure

$$\begin{aligned} D_{ij}^{(\eta)}[\mathbf{w}] &= \min_{\{t_k\}_1^n} \sum_{k=1}^{n} t_k d_{ijk} + \eta t_k \log\left(\frac{t_k}{w_k}\right), \qquad \sum_{k=1}^{n} t_k = 1, \\ &= -\eta \log\left\{\sum_{k=1}^{n} w_k \exp\left(-\frac{d_{ijk}}{\eta}\right)\right\}. \end{aligned} \tag{34}$$

This is a distance between objects (i, j) based on a weighted inverse exponential mean (24)–(25) of $\{d_{ijk}\}_{k=1}^{n}$ with scale parameter η.

As η becomes large, $D_{ij}^{(\eta)}[\mathbf{w}]$ approaches the ordinary distance (5),

$$\lim_{\eta\to\infty} (D_{ij}^{(\eta)}[\mathbf{w}]) = \sum_{k=1}^{n} w_k d_{ijk}. \tag{35}$$

Therefore, as the limit is approached this distance definition (34) can be used on the right-hand side of equation (30) or equation (33) to produce equivalent surrogate criteria for criterion (27).

For finite values of η alternative surrogate criteria are defined. These alternatives need not lead to equivalent surrogates for criterion (27) since they will not necessarily satisfy properties (31)–(32). However, setting the value of η in equation (34) to be the same as that used for λ in criterion (23) produces a criterion that is quite similar to criterion (23) when all weight values are taken to be equal, $\mathbf{W} = \{1/n\}$. This can be seen by first using equation (12) to express criterion (23) as

$$Q(c) = -\lambda \sum_{l=1}^{L} W_l \log\left[\frac{1}{n}\sum_{k=1}^{n}\left\{\prod_{\substack{c(i)=l\\c(j)=l}} \exp\left(-\frac{d_{ijk}}{\lambda}\right)\right\}^{1/N_l^2}\right]. \tag{36}$$

Setting $\eta = \lambda$ and then substituting equation (34) into equation (30) or (33) with all weight values equal to $1/n$ produces the surrogate criterion

$$\tilde{Q}(c) = -\lambda \sum_{l=1}^{L} W_l \log\left\{\prod_{\substack{c(i)=l\\c(j)=l}} \frac{1}{n}\sum_{k=1}^{n} \exp\left(-\frac{d_{ijk}}{\lambda}\right)\right\}^{1/N_l^2}. \tag{37}$$

Each term in both criterion (36) and criterion (37) contains the logarithm of a measure of central tendency of $\{\exp(-d_{ijk}/\lambda)\}$, for all $c(i) = c(j) = l$ and $1 \leqslant k \leqslant n$. For $Q(c)$ this measure is the arithmetic mean over k of the geometric mean over (i, j). For $\tilde{Q}(c)$ it is the geometric mean over (i, j) of the arithmetic mean over k. Both of these criteria are similar in that they are most strongly influenced by the d_{ijk} that have small values compared with λ, and correspondingly less influenced by those with larger values. By contrast, directly using definition (30) or (33) based on equation (17) using equal weight values $\mathbf{W} = \{1/n\}$ produces the criterion

$$\bar{Q}(c) = \sum_{l=1}^{L} \frac{W_l}{N_l^2} \sum_{c(i)=l} \sum_{c(j)=l} \frac{1}{n} \sum_{k=1}^{n} d_{ijk}. \tag{38}$$

Each term in equation (38) contains a measure of central tendency of $\{d_{ijk}\}$, for all $c(i) = c(j) = l$ and $1 \leqslant k \leqslant n$, based on the arithmetic mean. This criterion is independent of λ and most strongly influenced by the larger valued d_{ijk}. Therefore, to the extent that the respective geometric and arithmetic means of $\{\exp(-d_{ijk}/\lambda)\}$ appearing in criterion (36) and criterion (37) are not too different, solutions minimizing criterion (37) would probably be much closer to those minimizing criteria (23) and (36) than solutions produced by minimizing criterion (38). (Note that $0 < \exp(-d_{ijk}/\lambda) \leqslant 1$.) Empirical evidence suggests that this is indeed so.

Since $Q(c)$ and $\tilde{Q}(c)$ are not identical, applying a clustering algorithm based on definition (30) or definition (33), substituting equation (34) in place of equation (17) (with $\eta = \lambda$ and equal weights $\mathbf{W} = \{1/n\}$) does not produce the solution minimizing criterion (23). It only provides a potentially good starting-point for the iterative algorithm that is described in Section 5. From equation (35), as $\eta \to \infty$ this substitution produces the distance measure that is used by that algorithm. This suggests a homotopy optimization strategy in which equation (34) replaces equation (17) in definition (30) or definition (33), with η being the homotopy parameter. Its value is initialized to that of λ and then gradually increased as iterations proceed. This smoothly transforms the criterion being minimized from criterion (37), to criterion (27) based on definition (30) or definition (33), as the weight values progress from $\mathbf{W}=\{1/n\}$ to their minimizing values. Such a strategy leads to the following algorithm for clustering objects on subsets of attributes (COSA) (algorithm 1).

```
1: initialize—$\mathbf{W}=\{1/n\}$; $\eta=\lambda$.
2: loop {
3: compute distances $D_{ij}[\mathbf{W}]$ (30), (33) and (34)
4: $c \leftarrow$ clustering algorithm ($\{D_{ij}[\mathbf{W}]\}$)
5: compute weights $\mathbf{W}=\{\mathbf{w}_l\}_1^L$ (12) and (22)
6: $\eta=\eta+\alpha\lambda$
7: } until $\mathbf{W}$ stabilizes.
8: output—$c^*=c$.
```

For a given value of λ, the value of α (line 6) controls the rate of increase in the value of the homotopy parameter η. There is as yet no theory to suggest appropriate values of α in particular applications. Setting $\alpha=\infty$ causes this algorithm to compute the solution minimizing criterion (37) at the first iteration, and then immediately to switch to the algorithm that is described in Section 5, based on ordinary distance, expression (17), (30) or (33), starting at that solution. Smaller values of α cause a more gradual evolution from weighted inverse exponential distance (34) to ordinary distance as the weight values in turn evolve. Empirical evidence suggests that if the clustered groups tend to concentrate on small subsets of the attributes the value of α should be taken to be fairly small ($\alpha \lesssim 0.1$), causing a slow evolution. Otherwise, the weighted inverse exponential distance approaches ordinary distance too rapidly, thereby causing the algorithm to converge to an inferior local minimum in spite of its potentially good starting-point.

For inverse exponential distance (34) to approximate ordinary distance (35) closely, the value of the homotopy parameter η must become large compared with typical values of the interpoint distances d_{ijk} on each attribute k. As a consequence of their normalization (7) and (10) these attribute interpoint distances have expected values of 1 with typical values in the range $0 \lesssim d_{ijk} \lesssim 2$. For example, with normally distributed attribute values $\{x_{ik}\}$ and equal weights $\{w_k=1/n\}_1^n$, the correlation of the distances (34) for $\eta=1$ with those produced by $\eta=\infty$ is already 0.91, and for $\eta=2$ it is 0.97. This suggests that the transition of distance (34) to (35) is achieved when η reaches values in this range.

The algorithm, however, usually converges to equivalent solutions for much smaller values of η. This is caused by the weighting of the respective attributes within each clustered group after the first iteration (step 5). Attributes k with typically large interpoint distances ($d_{ijk} \gg \lambda$) receive small weights through equations (12) and (22) compared with those characterized by small interpoint distances ($d_{ijk} \lesssim \lambda$). Thus, both distance measures (34) and (35) are primarily influenced by those attributes k for which $d_{ijk} \lesssim \lambda$. This mechanism causes the transition of distance (34) to (35) to occur for much smaller values of η, typically $\eta \simeq \lambda$, when there is preferential clustering on attribute subsets. In fact, for all the applications that are presented in Section 12 below, using $0 \leqslant \alpha \leqslant 0.25$ caused the algorithm to converge to equivalent solutions, with some being invariant over a much broader range.

In applications where clustering tends to occur on relatively large numbers of attributes, larger values of α will tend to produce better results. However, it is in precisely these settings that the usual clustering algorithms (3) and (5) based on unweighted distances $\{w_k=1/n\}_1^n$ perform well, and little advantage is associated with the COSA strategy.

6.1. Scale parameter

The primary tuning parameter of the COSA procedure is the scale parameter λ (22). The goal is to identify groups of objects $G_l = \{i|c(i) = l\}$ (clusters) such that, on subsets of the attributes k, the characteristic interpoint distances d_{ijk} within each are relatively small $\{d_{ijk} \ll 1|c(i) = c(j) = l\}$. From equations (12) and (22) the value of λ defines the characteristic scale of these 'small' interpoint distances to which the procedure will have sensitivity. For large values $\lambda \gtrsim 1$ $(\eta \geqslant \lambda)$ both distance (34) and distance (35) reduce to ordinary distance with equal weights $\{w_{kl} = 1/n\}_1^n$ on all attributes within each cluster, so that COSA approximates ordinary clustering based on expressions (3) and (5). Thus if the goal is to uncover preferential clustering on subsets of attributes the value of λ should be taken to be small $(\lambda \ll 1)$.

As the value of λ is reduced, however, fewer objects within each group G_l have influence on the estimated weights through equations (12) and (22), thereby increasing the variance of these estimates and reducing the power of the procedure. Thus, λ can be regarded as a 'smoothing' parameter controlling a kind of bias–variance trade-off by analogy with more general density estimation procedures. Values that are too large give rise to oversmoothing, reducing sensitivity to narrow clustering on small subsets of attributes. Values of λ that are too small (undersmoothing) increase variance that also reduces power to uncover the overall clustering structure. Ideally, the value of λ should be set to the characteristic scale of the small distances d_{ijk} on those attributes k on which each of the groups G_l preferentially cluster. This is of course unknown. Variance considerations suggest somewhat larger values for smaller sample sizes.

Since an optimal value of λ is situation dependent and there is as yet no theory to suggest good values, the only recourse is to experiment with several values and to examine the results. Empirical evidence so far suggests that in the presence of sharp clustering on small subsets of attributes the procedure is usually not highly sensitive to values in the range $0.1 \leqslant \lambda \leqslant 0.4$. However, in the presence of more subtle structure the results can be fairly sensitive to a choice for its value.

7. Hierarchical clustering

The COSA algorithm 1 of the preceding section uses a conventional iterative clustering method as a primitive (line 4). It can be viewed as a 'wrapper' placed around a chosen clustering algorithm extending that algorithm to clustering on subsets of attributes. As with most conventional iterative clustering methods, the number of clusters sought L must be specified.

A very popular class of clustering techniques, especially with gene expression microarray data, are hierarchical methods. These do not require prespecification of the number of clusters. Instead, they arrange potential clusters in a hierarchy displayed as a binary tree ('dendrogram'). The user can then visualize this representation to assess the degree of clustering in the data and manually choose a particular partition of the objects into groups. Using the COSA algorithm 1 as a wrapper around such a manually driven procedure is cumbersome at best. For hierarchical clustering, we need a version of the algorithm that provides interobject distances $\{D_{ij}\}$ encouraging clustering on subsets of attributes, without requiring the specification of a particular iterative clustering algorithm or the number of groups L.

The key ingredient to producing such a version is based on the definition of clustering: pairs of objects (i, j) within the same solution-clustered group $c^*(i) = c^*(j)$, using a particular distance definition D_{ij}, will tend to have relatively small values of D_{ij}. This is the goal driving all clustering methods. Let KNN(i) be K closest objects to i based on D_{ij},

$$\mathrm{KNN}(i) = \{j|D_{ij} \leqslant d_{i(K)}\} \tag{39}$$

where $d_{i(K)}$ is the Kth order statistic of $\{D_{ij}\}_{j=1}^{N}$ sorted in ascending values. Then among those objects $j \in \mathrm{KNN}(i)$ there will be an overrepresentation of objects for which $c^*(i) = c^*(j)$, i.e.

$$\frac{1}{K} \sum_{j \in \mathrm{KNN}(i)} I\{c^*(j) = c^*(i)\} > \frac{1}{N} \sum_{j=1}^{N} I\{c^*(j) = c^*(i)\}. \tag{40}$$

The more pronounced the clustering, the stronger this inequality becomes. Therefore, to the extent that inequality (40) holds, statistics computed on $\mathrm{KNN}(i)$ will reflect those computed on $\{j | c^*(j) = c^*(i)\}$. In particular, for the scale measure (12) this implies

$$S_{k,c^*(i)} \backsimeq \frac{1}{K^2} \sum_{j \in \mathrm{KNN}(i)} \sum_{j' \in \mathrm{KNN}(i)} d_{jj'k}. \tag{41}$$

This represents a measurement of scale of the attribute x_k for objects $\{j | j \in \mathrm{KNN}(i)\}$. Furthermore, in the interest of reduced computation expression (41) can in turn be approximated by

$$S_{ki} = \frac{1}{K} \sum_{j \in \mathrm{KNN}(i)} d_{ijk}. \tag{42}$$

Under these assumptions we can modify the COSA algorithm 1 by replacing the clustering algorithm (line 4) by a procedure that computes $\{\mathrm{KNN}(i)\}_1^N$, and replacing $\mathbf{w}_{c(i|\mathbf{W})} \leftarrow \mathbf{w}_i = \{w_{ki}\}_{k=1}^{n}$ in equations (30), (33) and (34) for computing the distances (line 3), with

$$w_{ki} = \exp\left(-\frac{S_{ki}}{\lambda}\right) \Big/ \sum_{k'=1}^{n} \exp\left(-\frac{S_{k'i}}{\lambda}\right) \tag{43}$$

for calculating the weights (line 5). With this substitution, the matrix of weights $\mathbf{W}$ becomes an $n \times N$ matrix with entries w_{ki}. These changes produce the following COSA algorithm 2.

```
1: initialize—$\mathbf{W} = \{1/n\}$; $\eta = \lambda$.
2: loop {
3: compute distances $D_{ij}[\mathbf{W}]$ (30), (33) and (34)
4: compute $\{\mathrm{KNN}(i)\}_1^N$ (39)
5: compute weights $\mathbf{W} = \{w_{ki}\}$ (42) and (43)
6: $\eta = \eta + \alpha\lambda$
7: } until $\mathbf{W}$ stabilizes
8: output—$\{D_{ij} = D_{ij}[\mathbf{W}]\}$.
```

The purpose of this algorithm is to obtain a good set of weight values $\mathbf{W} \in \mathbf{R}^{n \times N}$ for calculating interpoint distances $\{D_{ij}\}$ (step 8) by approximately minimizing the criterion

$$Q(\mathbf{W}) = \sum_{i=1}^{N} \left\{ \frac{1}{K} \sum_{j \in \mathrm{KNN}(i)} D_{ij}[\mathbf{w}_i] + \lambda \sum_{k=1}^{n} w_{ki} \log(w_{ki}) \right\}. \tag{44}$$

These distances can then be input to hierarchical clustering algorithms.

The weight values $\mathbf{W}^* = \{\mathbf{w}_i^*\}_1^N$ minimizing criterion (44) are those that create the smallest K nearest neighbourhoods, subject to the negative entropy incentive (19). Here the size of each neighbourhood is measured by the average distance to its centre point $\mathbf{x}_i$, using attribute weights $\mathbf{w}_i = \{w_{ki}\}_1^n$. This is inversely related to an estimate, based on $\mathrm{KNN}(i)$, of the probability density $p(\mathbf{x}_i | \mathbf{w}_i)$. In this sense, the solution weights minimizing criterion (44) are chosen to maximize these probability density estimates.

The considerations concerning the value of the scale parameter λ and homotopy rate parameter α (line 6) are the same as those for the COSA algorithm 1 that was discussed in Section 6 above. The size K that is chosen for the nearest neighbourhoods is not critical and results are fairly stable over a wide range of values. It should be sufficiently large to provide stable estimates of S_{ki} (42) but not too much larger than the size of the cluster containing the ith object. Setting $K \backsimeq \sqrt{N}$ is a reasonable choice, although some experimentation may be desirable after reviewing the sizes of the clusters uncovered.

8. Robust dispersion measures

Measurements of the dispersion of attribute values for sets of objects (10), (12) and (42) play an important role in the COSA procedures. These dispersion measures are based on computing mean values of the interobject distances on the respective attributes. For numeric-valued attributes (8), mean statistics are known to be highly sensitive to a small number of objects with unusually large values ('outliers'). Using medians as an alternative measure of central tendency eliminates this problem, making the overall procedure more robust. Since robustness is an important property for any data mining method, we replace the respective mean values in equations (10), (12) and (42) with medians for numeric-valued attributes. Furthermore, for computational efficiency, equation (10) is approximated by

$$s_k \simeq \mathrm{IQR}(\{x_{ik}\}_{i=1}^{N})/1.35 \tag{45}$$

where IQR is the interquartile range. The divisor in approximation (45) is the value that is appropriate for a normally distributed variate. (The corresponding values for a uniform and log-normal distributions are 1.12 and 1.62 respectively, so 1.35 represents an average choice.) For categorical (nominal) attributes there is no corresponding outlier issue so equations (10), (12) and (42) can be used to compute the respective attribute dispersions.

9. Interpretation

If the clustering procedure is successful in uncovering distinct groups (clusters), we would like to know whether each such group represents clustering on a subset of the attributes and, if so, to identify the relevant attribute subsets for each of the respective groups. Clustering algorithms used with COSA only report group membership (1). With this approach, there is no *explicit* attribute subset selection. However, the relative importance (relevance) of each attribute k to the clustering of each clustered group G_l is given by equation (22) substituting the robust (Section 8) analogue of S_{kl} (12)

$$S_{kl} = \frac{1}{N_l} \sum_{i \in G_l} \mathrm{median}\{d_{ii'k}\}_{i' \in G_l} \tag{46}$$

for numeric-valued attributes.

An unnormalized first-order approximation

$$I_{kl} = (S_{kl} + \varepsilon)^{-1} \tag{47}$$

can be interpreted as an *absolute* measure of the importance I_{kl} of attribute k to the clustering of solution group G_l. Here, ε^{-1} represents the maximum obtainable importance value. When computed over all objects in the data set, rather than over objects within an individual cluster, equation (47) evaluates to $I_k = (1+\varepsilon)^{-1}$ for all attributes, owing to the normalization in equations (7) and (10). Thus we can interpret equation (47) as inversely measuring the spread of

x_{ik}-values within the group G_l relative to its corresponding spread over all objects. For example, a value of $I_{kl} = 4$ implies that the spread of x_{ik}-values within G_l is roughly a quarter of that over all the objects in the data set. Large values of I_{kl} indicate that G_l is highly clustered on attribute x_k, whereas small values indicate the opposite. Inspection of the values of $\{I_{kl}\}_{k=1}^{n}$ for each cluster G_l allows us to ascertain the relevant attributes contributing to the clustering of the lth group G_l. Illustrations are provided in Section 12.

10. Missing values

In many applications there are incomplete data; some of the attribute values for the objects are missing. The distance measure (34) can be modified to accommodate missing values while taking advantage of the information that is present in the non-missing values. We simply make the modification

$$w_k \leftarrow w_k \; I(x_{ik} \neq \text{missing}) \; I(x_{jk} \neq \text{missing}) \tag{48}$$

in equation (34), and then renormalize the weights to sum to 1. This assigns a weight value of 0 to the kth attribute in the distance calculation if its value is missing on either object i or object j. If the two objects have no non-missing values in common, they are assigned an infinite distance so that they will not be placed in the same cluster.

For the calculation of the weights (22) and (43) on the kth attribute, only non-missing values of that attribute (x_k) are used to calculate S_{kl} (12) or S_{ki} (42). If in equation (42) object i is missing a value for x_k, or all K nearest neighbours of object i are missing values of x_k, then the corresponding weight is set to 0: $w_{ki} = 0$.

11. Targeted clustering

The COSA algorithms attempt to uncover distinct groups of objects that have similar joint values on subsets of the attributes. The actual joint values of the attributes in the subset about which the objects cluster is unspecified; the attempt is to find clustering centred on any possible joint values of the attributes. This may not always be the goal; there may be preferred values on some or all of the attributes about which we would like to focus.

For example, we might have data on the spending habits of consumers in terms of amounts spent on various products or activities. The goal might be to identify groups of consumers (objects) who spend relatively large amounts on subsets of the products (attributes), and to be unconcerned with those who spend moderate to small amounts. Attempting to find arbitrary clustering could obscure small but potentially interesting clusters of such high spenders. Alternatively, we might be interested in identifying clusters of low spenders, or perhaps clusters of extreme spenders who either spend excessively large or small but not moderate amounts on various items. In contrast, specific consumer research might want to focus on consumers who do in fact spend moderate amounts of money. Similarly, in gene expression data we might seek clusters of samples (objects) that have preferentially high or low or extreme expression levels on subsets of the genes (attributes). Again seeking clusters centred at arbitrary values can cause difficulty in uncovering the structure of interest, especially if it is fairly subtle.

11.1. Single-target clustering

Focused or targeted clustering can be accomplished by modifying the distance definitions (7) on selected individual attributes. Let t_k be a predefined target value on the kth attribute and

(x_{ik}, x_{jk}) be the corresponding respective values of objects i and j on that attribute. Define the 'targeted' distance between objects (i, j) on the kth attribute as

$$d_{ijk}(t_k) = \max\{d_k(x_{ik}, t_k), d_k(x_{jk}, t_k)\}, \tag{49}$$

where

$$d_k(x, t) = |x - t|/s_k, \tag{50}$$

with s_k given by approximation (45) for numeric attributes, and

$$d_k(x, t) = I(x \neq t)/s_k \tag{51}$$

with s_k given by equation (10) for categorical (nominal) attributes. This distance (49)–(51) is small only if both the values of x_{ik} and x_{jk} are close to each other *and* close to the target value t_k. Using expressions (49)–(51) in place of (7)–(9) for any attribute or set of attributes will cause the clustering algorithm, when considering groupings on those attributes, to consider only clusters near the targeted values. This can substantially reduce the cluster search space, making subtle clustering near the target values easier to uncover.

For the consumer spending data example, setting target values near the maximum data value on each attribute will cause a clustering algorithm to seek only clusters of high spenders. Similarly clusters of only high (or low) gene expressions can be sought through the same mechanism. In both cases restricting the search makes it more likely to find the targeted clusters of interest, since the algorithm will not be distracted by other perhaps more dominant (but less interesting) clustering.

11.2. Dual-target clustering

Single-target clustering (49)–(51) can be quite powerful in uncovering subtle clustering effects as will be illustrated in Section 12.1. However, for some applications it can be too restrictive. In the consumer spending problem we may be interested in clusters of 'extreme' spenders, people who either spend unusually high or low amounts on sets of items. Similarly, we might be seeking clusters of samples with unusually high or low (but not moderate) gene expression levels. This type of clustering can be accomplished by using 'dual-target' distances

$$d_{ijk}(t_k, u_k) = \min\{d_{ijk}(t_k), d_{ijk}(u_k)\} \tag{52}$$

on selected attributes x_k, where $d_{ijk}(\cdot)$ is the corresponding single-target distance (49). This distance (52) is small whenever x_{ik} and x_{jk} are either both close to t_k *or* both close to u_k. In the consumer spending and gene expression examples we might set t_k and u_k respectively to values near the maximum and minimum data values of the attributes. Using definition (52) with COSA will cause the clustering algorithm to seek clusters of extreme attribute values, ignoring (perhaps dominant) clusters with moderate attribute values.

12. Illustrations

In the following sections we illustrate COSA on several data sets. In all the examples that are presented here the COSA algorithm 2 (Section 7) was employed with average linkage hierarchical clustering so that the resulting cluster structures can be visualized. The value of the scale parameter λ (23) was taken to be $\lambda = 0.2$ and the number of nearest neighbours K (39) was taken to be the square root of the sample size. For the attribute importance calculations (47), ε was set to 0.05.

12.1. Simulated data

In this section we present a modest systematic investigation of the properties of COSA-based clustering, and its relationship to traditional approaches based on Euclidean distance. Both are applied to a series of simulated data sets of sizes that are characteristic of those produced by gene expression microarray experiments. Specifically all data sets consisted of $N = 100$ objects and $n = 10000$ attributes. To aid interpretation, the simulated clustering structure was taken to be very simple.

Each data set consisted of two groups (clusters). The first group was a random sample of 85 objects drawn from a 10000-dimensional standard normal distribution. The second group of 15 objects was also drawn from a 10000-dimensional normal distribution, but its first n_0 attributes each had a mean of $\mu = 1.5$ and standard deviation $\sigma = 0.2$. The remaining $10000 - n_0$ attributes of the second group each had zero mean and unit standard deviation. Thus, the population distributions of the two groups differ only on the first n_0 attributes. After generation, the pooled sample was standardized to have zero mean and unit variance on all attributes. These

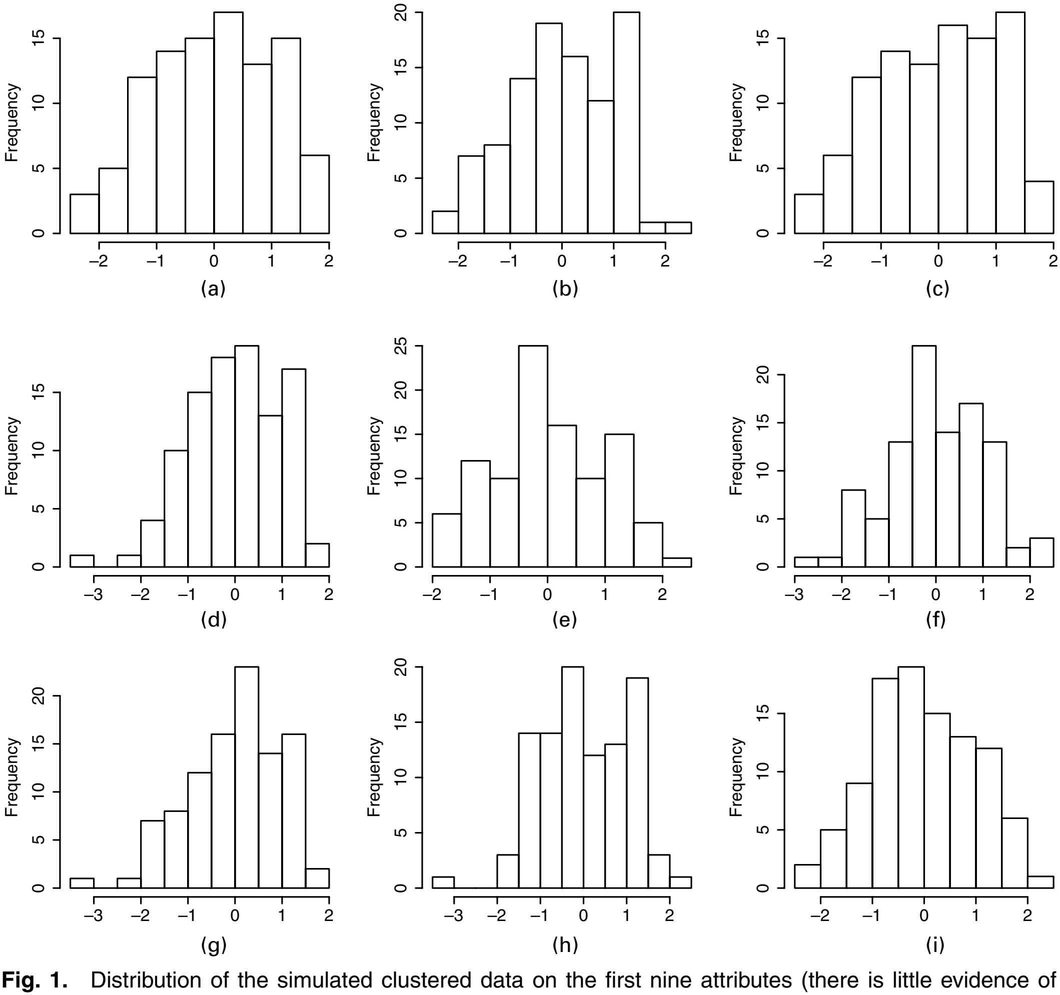

Fig. 1. Distribution of the simulated clustered data on the first nine attributes (there is little evidence of obvious clustering in these marginal distributions): (a) variable 1; (b) variable 2; (c) variable 3; (d) variable 4; (e) variable 5; (f) variable 6; (g) variable 7; (h) variable 8; (i) variable 9

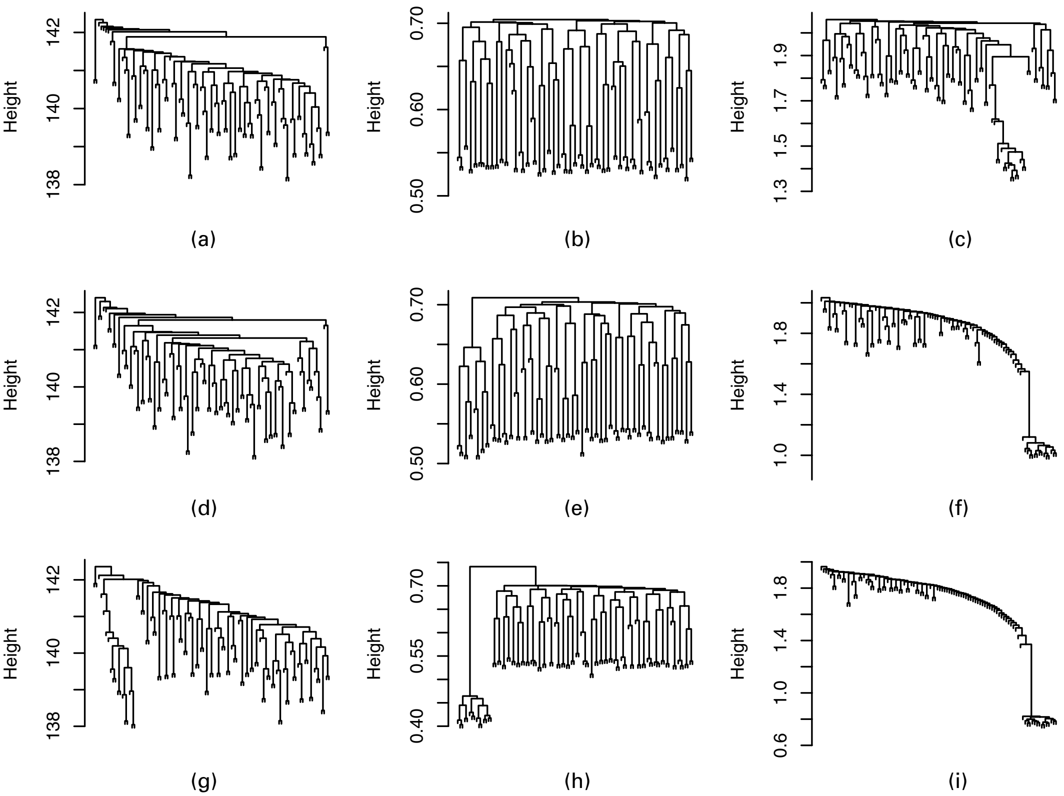

Fig. 2. Average linkage dendrograms for three simulated data sets of 100 objects with 10000 attributes (each data set consists of a small 15-object group clustering on n_0 attributes, nested within an unclustered background of 85 objects; Euclidean distance is seen to require clustering on more attributes than are required by the COSA distances to detect the smaller group; targeted COSA provides the most power in this setting): (a) Euclidean distances, $n_0 = 10$; (b) COSA distances, $n_0 = 10$; (c) targeted COSA distances, $n_0 = 10$; (d) Euclidean distances, $n_0 = 60$; (e) COSA distances, $n_0 = 60$; (f) targeted COSA distances, $n_0 = 60$; (g) Euclidean distances, $n_0 = 150$; (h) COSA distances, $n_0 = 150$; (i) targeted COSA distances, $n_0 = 150$

data thus contain a small group that exhibits clustering on only a few (n_0) attributes, together with a large non-clustered background. The purpose is to study the ability of the respective clustering approaches to uncover the second small group as a function of the number of attributes n_0 on which it clusters. Fig. 1 shows histograms of the pooled data on the first nine attributes ($n_0 > 9$). Histograms of the other $n_0 - 9$ clustered attributes are similar in that they show little evidence of clustering on any of the individual marginal distributions of the attributes that are relevant to the clustering.

Clusterings based on three distance measures are compared: the squared Euclidean distance, non-targeted COSA distance and single-target COSA distance with the target on each attribute set to the 95th percentile of its data distribution. Fig. 2 shows average linkage dendrograms for three values of n_0 ($n_0 = 10$ (Figs 2(a)–2(c)), $n_0 = 60$ (Figs 2(d)–2(f)) and $n_0 = 150$ (Figs 2(g)–2(i))) for each of the three distance measures (Euclidean (Figs 2(a), 2(d) and 2(g)), non-targeted COSA (Figs 2(b), 2(e) and 2(h)) and single-target COSA (Figs 2(c), 2(f) and 2(i)).

For $n_0 = 10$, clustering based on targeted COSA distance readily distinguishes the small 15-object cluster from the background; Euclidean and non-targeted COSA distances cannot do so. At $n_0 = 60$, targeted COSA dramatically distinguishes the small group, whereas non-targeted COSA is seen to be barely able to provide separation (extreme left). Euclidean distance still shows no evidence of the smaller group. With $n_0 = 150$, Euclidean distance begins to provide evidence of the smaller group, whereas both COSA distances clearly delineate it.

Although simple in structure, this example provides some insight into the relative strengths of the three distance measures. It illustrates the ability of COSA-based procedures to separate small groups clustering only on a tiny fraction of the total number of attributes. Non-targeted COSA could detect clustering on less than half the number of attributes required by Euclidean distance. Targeting, when appropriate, is seen to be especially powerful. It could detect a cluster involving only 15% of the objects and 0.1% of the attributes.

This example is especially simple in that the only structure is the existence of the small cluster; the larger distances reflect pure unstructured noise. In actual problems this is seldom so. Larger distances on the attributes are likely to exhibit considerable structure that may or may not be associated with clustering. Since Euclidean distance is especially sensitive to the larger distances on individual attributes, this large scale structure can obscure the detection of smaller groups clustering on subsets of the attributes. COSA is designed to be especially sensitive to small distances on the attributes, thereby being more sensitive to this type of structure.

The upper (dark) curve in Fig. 3 shows the 100 largest attribute importance values (47) evaluated on the 15-object group detected with targeted COSA on the simulated data for $n_0 = 10$ (Fig. 2(c)). The importance values for the first 10 (most important) attributes are seen to be sharply higher than those for the other attributes. Their importance values range from 3.4 to 6.0 with a mean of 4.8. The population values are all 5.0 ($\sigma = 0.2$). The 10 lower (light) curves represent the 100 highest attribute importance values for 10 groups, each of 15 objects *randomly* selected from the data; the central curve is their average. The importance spectrum of the actual clustered group closely coincides with those of the randomly selected groups except for the 10 highest importance values. These 10 attributes that are estimated to be the most important

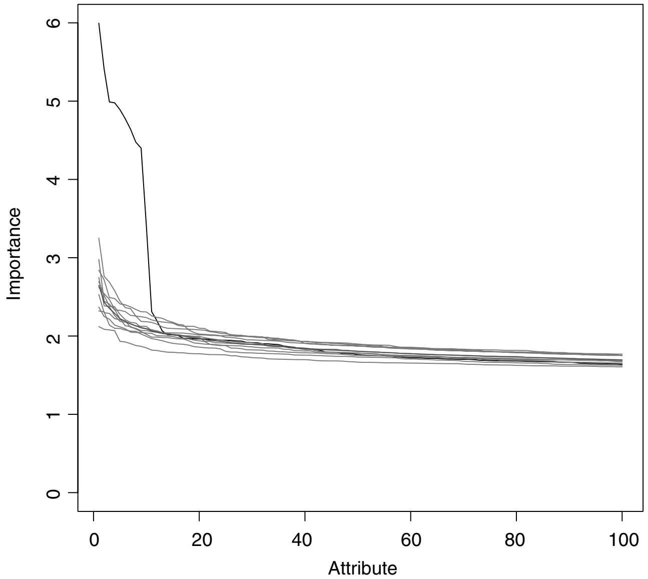

Fig. 3. 100 largest attribute importance values for the 15-object group detected by targeted COSA on the simulated data for $n_0 = 10$ (———), and the 100 highest attribute important values for 10 groups, each of 15 objects *randomly* selected from the data (———): the central curve is the average of these 10 curves; the detected group shows strong evidence of clustering on 10 attributes, with little or no evidence of clustering on the remaining 9990 attributes

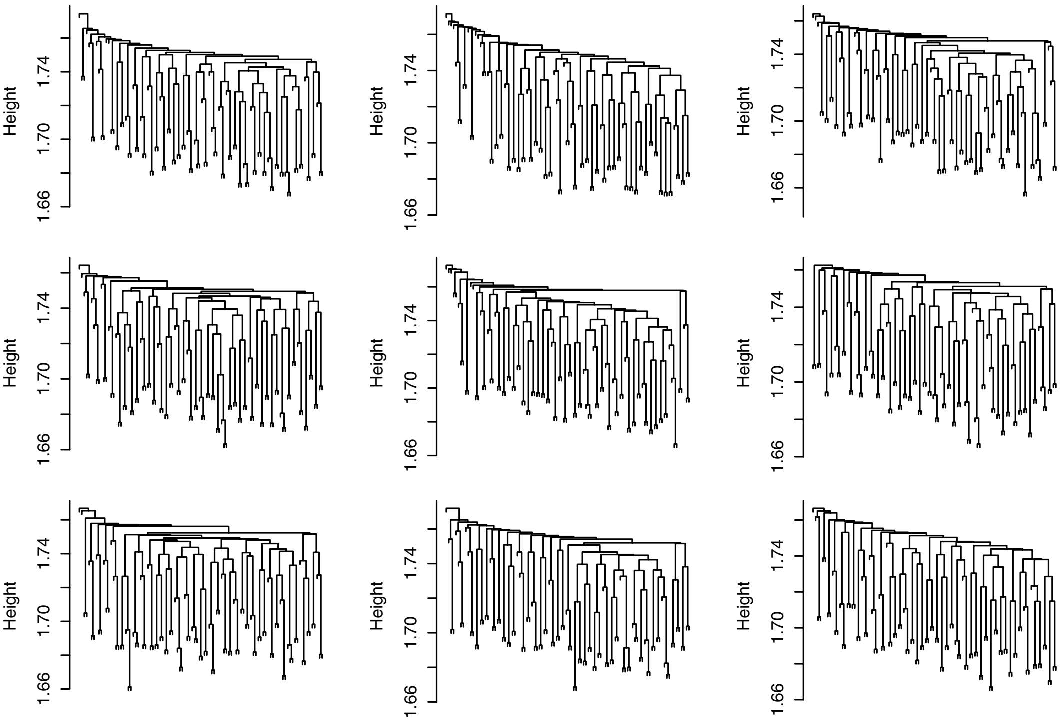

Fig. 4. Average linkage dendrograms resulting from applying single-target COSA distance to nine different data sets of 100 objects randomly drawn from a 10 000-dimensional standard normal distribution: no indications of obvious clustering appear in any of these plots

turn out to be $x_1, \ldots, x_{10}$, those actually relevant to forming the small cluster in the generating population.

This example demonstrates the sensitivity of COSA-based clustering in uncovering small groups that cluster only on very small subsets of the attributes. A potential worry is that such groupings may exist with sufficient strength in random data, sampled from an unstructured population, so that they will also be detected by COSA. The examples that are shown in Fig. 4 indicate that this is not likely. Shown are nine average linkage dendrograms resulting from applying single-target COSA to nine different data sets of 100 objects, each randomly sampled from a 10 000-dimensional standard normal distribution. As can be seen, there are no obvious indications of clustering in these plots. The corresponding dendrograms based on dual-target and non-targeted COSA (not shown) are similar in that they also show no obvious clustering.

12.2. Mitochondrial ribonucleic acid relative abundance data

The mitochondrial ribonucleic acid (RNA) data set consists of $n = 6221$ mitochondrial RNA relative abundance estimates derived from gene expression data (attributes), with $N = 213$ samples (objects). The data are an agglomeration of samples from 12 experiments derived from nine studies (Aach *et al.*, 2000). Attributes with more than half of their values missing were deleted from the analysis, leaving 6141 attributes still containing many missing values.

Fig. 5 displays the average linkage dendrogram that is obtained from the squared Euclidean distance on the standardized attributes. Substantial clustering is apparent. Five distinct groups that each contain 10 or more objects can be identified. These are delineated in Fig. 6(a). Fig. 7 shows the corresponding dendrogram based on (standard) COSA distance. With COSA, the

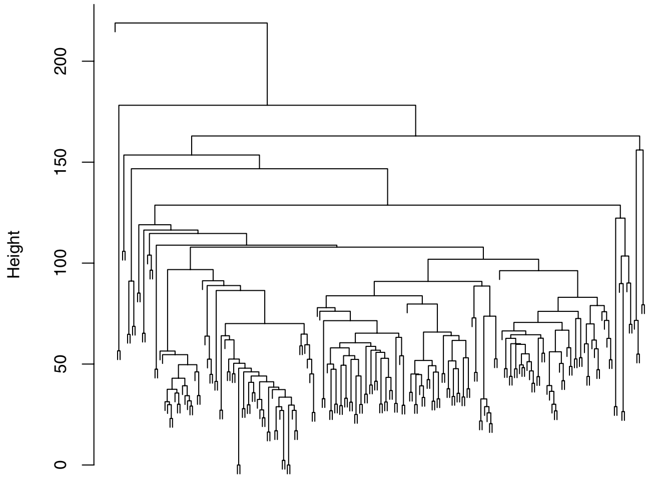

Fig. 5. Average linkage dendrogram based on Euclidean distance for the yeast mitochondrial RNA relative abundance data: substantial clustering is indicated

Table 1. Experiments comprising each of the Euclidean distance clusters for the yeast mitochondrial RNA relative abundance data

Cluster	1	2	3	4	5
Experiment	Cho	Hol	Spe_alpha	Spe_elut	Spe_cdc

Table 2. Experiments comprising each of the COSA distance clusters for the yeast mitochondrial RNA relative abundance data

Cluster	1	2	3	4	5	6	7	8	9
Experiment	Hol	Hol	Cho	Spe_cdc	Spe_cdc	Spe_elut	Spe_alpha	Chu	Der_duix (+)

separation into distinct clusters is seen to be much sharper, and at least nine distinct groups (containing more than 10 objects) can be identified. These are delineated in Fig. 6(b).

Each of the Euclidean clusters (Fig. 6(a)) uniquely contain all the objects (samples) arising from five of the 12 experiments. These are identified in Table 1 with the delineated clusters labelled sequentially from left to right. Unsupervised Euclidean distance clustering could separate these five experiments from the rest of the data in the absence of an experiment label.

Eight of the nine clusters that were identified in the COSA dendrogram (Fig. 6(b)) contain objects (samples) from unique experiments. These are identified in Table 2 with the COSA clusters in Fig. 6 labelled sequentially from left to right. COSA clusters 3, 6, 7 and 8 contain all the objects from each of the corresponding experiments. Clusters 1 and 2 partition all the Hol experimental samples into two distinct groups, whereas clusters 4 and 5 similarly divide all the samples of the Spe_cdc experiment. Cluster 9 is the only impure cluster, containing all the Der_duix samples and a few samples from other experiments as well.

The results from COSA clustering suggest that the Hol and Spe_cdc experiments each partition into two distinct groups of similar size. This is not evident from Euclidean-distance-based clustering.

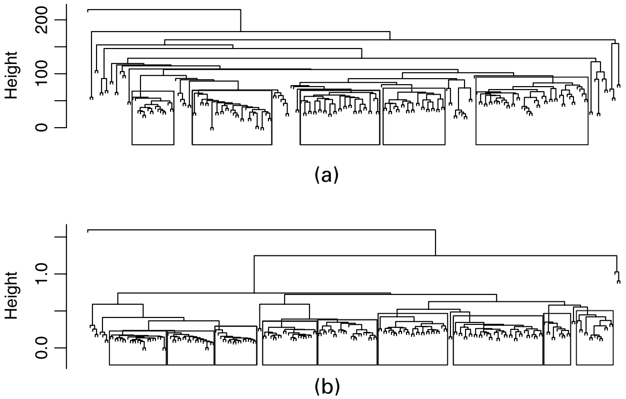

Fig. 6. Average linkage dendrograms based on (a) Euclidean and (b) COSA distance for the yeast mitochondrial RNA relative abundance data, with clustered groups involving 10 or more objects delineated

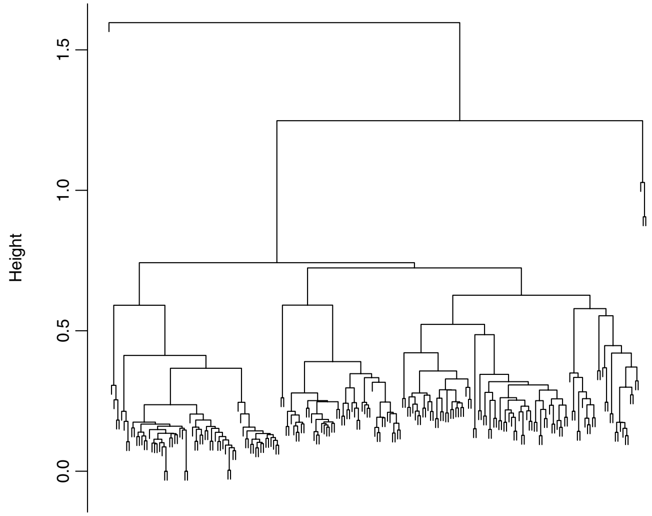

Fig. 7. Average linkage dendrogram based on COSA distance for the yeast mitochondrial RNA relative abundance data: very sharp clustering is apparent

Fig. 8 illustrates the attribute importance values (upper dark curves) for the two Hol groups. The lower (light) curves are the corresponding ordered attribute importances for same-sized groups randomly selected from the whole data set. Both of the Hol subgroups strongly cluster on a relatively small fraction of all the attributes. The concentration is somewhat sharper for the first (left-hand) group. The attribute subsets on which the two groups strongly cluster are not identical but substantially overlap. There were 41 common attributes among the 100 most relevant for each group.

Euclidean-distance-based clustering could partition five of the six experiments that contain 10 or more samples (objects) into separate groups. (The other six experiments contained fewer than 10 samples.) COSA clustering (more sharply) separated all six of these experiments (with a contaminated seventh cluster) and in addition could detect strong clustering structure *within* two of them.

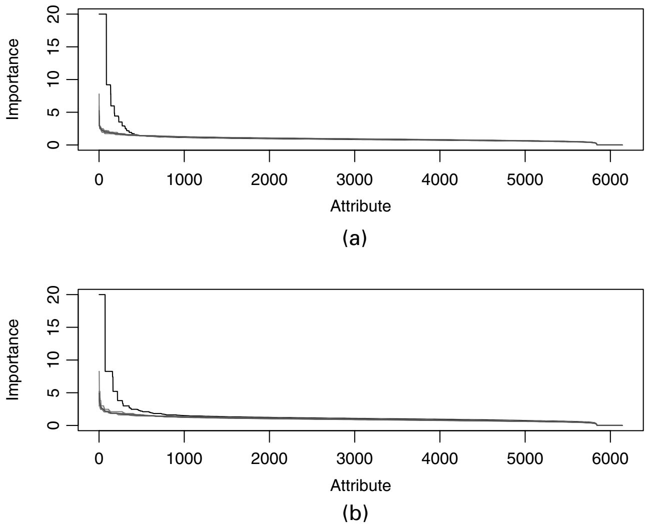

Fig. 8. Ordered attribute importances (———) for (a) Hol group 1 (23 objects) and (b) Hol group 2 (19 objects) (both of these groups exhibit strong clustering only on a small subset of the attributes (genes): ———, corresponding ordered importances for randomly selected groups of the same size (the central curve is their average)

12.3. Medical data

The medical data were collected at the Leiden Cytology and Pathology Laboratory (see Meulman *et al.* (1992)). They consist of $n = 11$ manually observed features (attributes) of cells taken from pap smears of $N = 242$ patients (objects) with cervical cancer or its precursor lesions (dysplasias). Attributes $(x_1, \ldots, x_4)$ and (x_6, x_7, x_8) are abnormality ratings by a pathologist of various aspects of each cell. The ratings range from 1 (normal) to 4 (very abnormal). Most of these attributes have only three distinct values (normal ratings were rare) so they were treated as being categorical. The remaining four features $(x_5, x_9, x_{10}, x_{11})$ are numerical (counts) with many distinct values.

The strongest clustering was revealed by using dual-target distance (52), with the targets set to the 5th and 95th percentiles of the data distribution on each numeric attribute. No targets were specified for the categorical attributes. Fig. 9 shows the resulting average linkage dendrogram. These data are seen to partition into nine fairly distinct groups, containing 10 or more objects, delineated by the corresponding rectangles. Moderate additional clustering within some of these groups is also indicated.

The attribute importances (47) for each of these nine groups are plotted (on a square-root scale) in Fig. 10. The groups are displayed in their dendrogram (Fig. 9, from left to right) order. All the groups exhibit very strong clustering on 1–3 attributes, with some groups showing moderately strong clustering on a few other attributes as well. Each group is seen to cluster on a different small subset of the attributes, with some overlap between the subsets.

In addition to the 11 cell features that were taken from their pap smears, each patient was diagnosed in a subsequently performed biopsy. Each of these histological diagnoses was assigned a numerical score, with values ranging from 1 to 5, reflecting the severity of the dysplasia (mild, moderate or severe) or cervical carcinoma (*in situ* or invasive). Fig. 11 shows the distribution (box plot) of these score values for patients who were assigned to each of the nine clustered groups that were identified in Fig. 9 (from left to right) and shown in Fig. 10. The labels along the abscissa show the median of the index values within each of the respective groups. Each

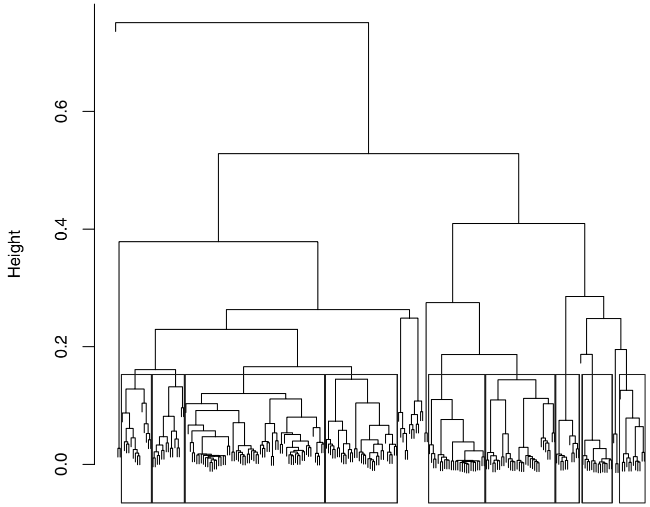

Fig. 9. COSA average linkage dendrogram for the medical data using high–low dual-target dissimilarities on the numeric variables: these data are seen to partition into nine fairly well-separated groups of more than 10 objects each as delineated by the rectangles

of the uncovered clusters, based only on the cell features, is seen to correspond to a relatively narrow range of increasing score values, indicating a substantial relationship between the COSA group membership and the severity of the diagnosis, with a clear separation between carcinoma (severity scores 4 and 5) and dysplasia (severity scores 1, 2 and 3).

13. Discussion

COSA can be viewed as an enhancement to distance-based clustering methods, enabling them to uncover groups of objects that have preferentially close values on different, possibly overlapping, subsets of the attributes. There do not appear to be other distance-based methods that are directly focused on this goal. There are, however, non-distance-based modelling methods that have been proposed for this purpose.

The one closest in spirit is product density mixture modelling (AutoClass—Cheesman and Stutz (1996); see also Banfield and Raftery (1993)). The joint distribution of the attribute values is modelled by a mixture of parameterized component densities. Each component in the mixture is taken to be a product of individual probability densities on each of the attributes. Prior probability distributions are placed on all model parameter values and a heuristic search strategy is used to attempt to maximize the posterior probability on the data. Each of the components in the resulting solution is considered to be a ‘soft’ cluster.

As with COSA, the (posterior probability) criteria being optimized by these methods are highly non-convex functions of their parameters and avoiding convergence to distinctly inferior local optima is a problem (see Ghosh and Chinnaiyan (2002)). Furthermore for large data sets, such as those derived from gene expression microarrays, the very large number of associated model parameters causes severe computational and statistical estimation difficulties. Therefore, specialized preprocessing and screening procedures are required to reduce the size of the problem substantially to manageable proportions. Also, experimenting with various data transformations is often required in an attempt to bring the data into conformity with the parametric model (Yeung *et al.*, 2001). COSA is distinguished from these methods by its

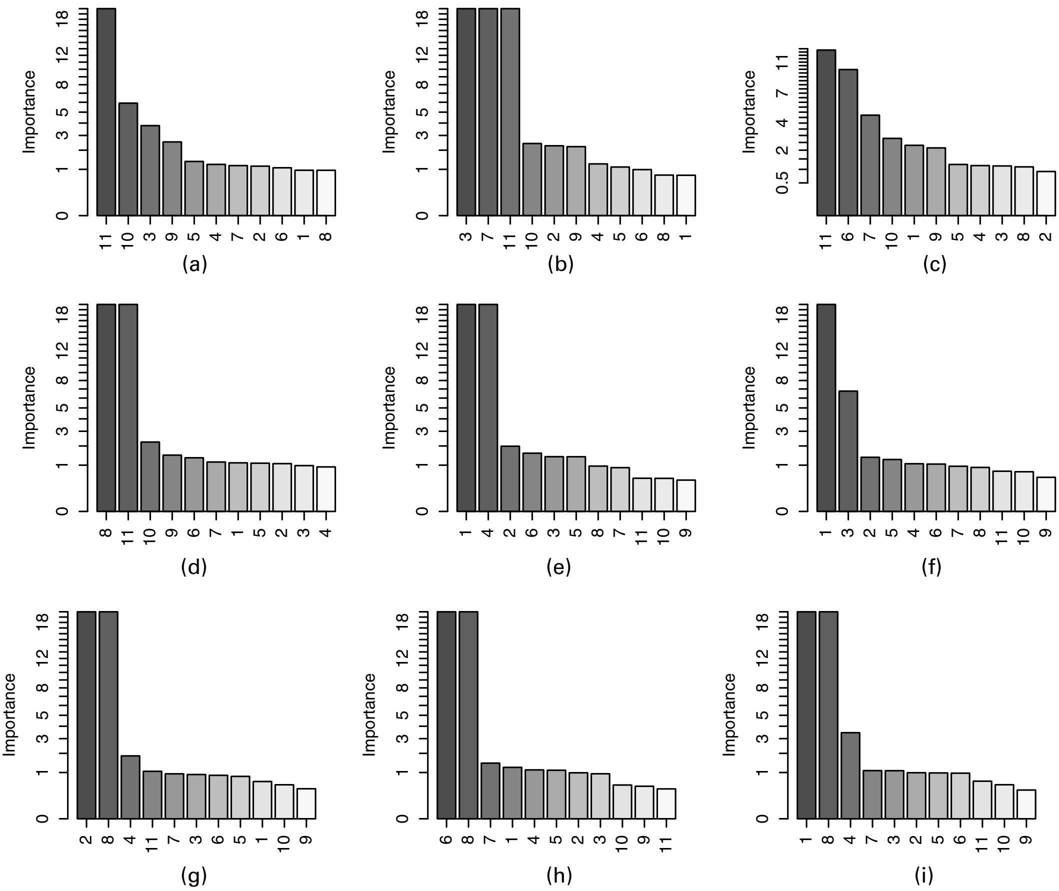

Fig. 10. Attribute importances for each of the nine groups that were uncovered in the medical data set, shown on a square-root scale—each of these groups tends to cluster on a relatively small subset of the 11 attributes: (a) group 1, 12 objects; (b) group 2, 28 objects; (c) group 3, 94 objects; (d) group 4, 10 objects; (e) group 5, 13 objects; (f) group 6, 37 objects; (g) group 7, 18 objects; (h) group 8, 15 objects; (i) group 9, 14 objects

nonparametric formulation and computational feasibility on large data sets, thereby reducing or eliminating dependence on customized preprocessing and screening procedures. It can be used with hierarchical clustering methods, and it employs a search strategy using a particular homotopy technique in an attempt to avoid distinctly suboptimal solutions.

Motivated by gene expression microarray data, several recent techniques have been proposed to uncover clustering by directly modelling the (numeric) data matrix $\mathbf{X}=[x_{ij}]\in R^{N\times n}$ by additive decompositions. Each additive term is interpreted as a cluster. Plaid models (Lazzeroni and Owen, 2000) treat the objects and attributes symmetrically. The data matrix is represented by an expansion that is analogous to the singular value decomposition. The components of the singular vectors for each term ('layer') in the expansion are restricted to the two values $\{0,1\}$. A value 1 or 0 for the ith component of a left singular vector indicates that the corresponding ith row of the data matrix respectively does or does not contribute to the clustering that is represented by that layer. Similarly, a value 1 or 0 for the jth component of a right singular vector indicates that the corresponding column respectively does or does not contribute. Each layer

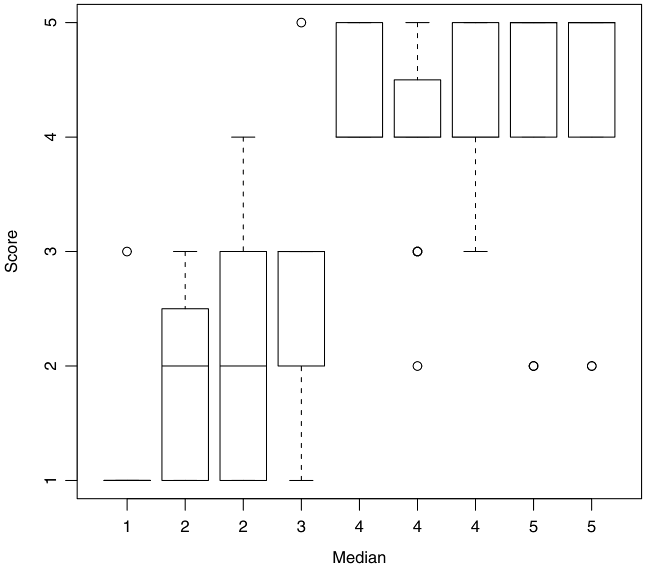

Fig. 11. Distribution of diagnosis indices (from 1, mild dysplasia, to 5, invasive carcinoma) within each of the nine clusters delineated (from left to right) in Fig. 9: the median index value for each group is shown along the abscissa; each group corresponds to a relatively narrow range of increasing index values, indicating a relationship between cluster membership and severity of disease

is interpreted as modelling the data matrix after subtracting the contributions of all previous layers. Gene shaving (Hastie *et al.*, 2000) seeks to decompose the $N \times n$ data matrix into a set of smaller $N_k \times n$ matrices ($1 \leqslant k \leqslant K$; $N_k \ll N$) such that within each the components of the row mean vector (averaged over the columns) exhibit high variance. The rows within each such matrix are interpreted as clusters.

Although by no means the same, the underlying goals of all these methods are quite similar. As with all such methods, a major component is the particular heuristic search strategy that is employed. Even with similar (or the same) goals, different search strategies have the potential to reach quite different solutions, representing different clustering structures, many of which may be interesting and useful. The particular characteristics of the COSA method that is proposed in this paper include the ‘crisp’ or ‘hard’ clustering of objects on possibly overlapping subsets of attributes, the use of targets anywhere within the domain of each attribute to focus the search on particular types of ‘interesting’ structure and, as noted above, a homotopy technique based on weighted inverse exponential distance to avoid suboptimal solutions. We conjecture that the use of this technique is crucial in finding the weights for the attributes that define the subsets for each separate cluster of objects. The COSA technique can be used in conjunction with a wide variety of (distance-based) clustering algorithms, including hierarchical methods, each employing its own particular encoder search strategy. As with any data analytic procedure, the validity and usefulness of the output of different clustering methods can only be evaluated by the user in the context of each particular application.

Acknowledgements

The work of Jerome H. Friedman was partially supported by the Department of Energy under contract DE-AC03-76SF00515 and by the National Science Foundation under grant DMS-97-64431.

References

Aach, J., Rindone, W. and Church, G. M. (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.*, **10**, 431–444.

Arabie, P., Hubert, L. J. and De Soete, G. (eds) (1996) *Clustering and Classification*. River Edge: World Scientific.

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Brusco, M. J. and Cradit, J. D. (2001) A variable selection heuristic for K-means clustering. *Psychometrika*, **66**, 249–270.

Cheesman, P. and Stutz, J. (1996) Bayesian classification (AutoClass): theory and results. In *Advances in Knowledge Discovery and Data Mining* (eds U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy), pp. 153–180. Cambridge: MIT Press.

DeSarbo, W. S., Carroll, J. D., Clarck, L. A. and Green, P. E. (1984) Synthesized clustering: a method for amalgamating clustering bases with differential weighting of variables. *Psychometrika*, **49**, 57–78.

De Soete, G. (1986) Optimal variable weighting for ultrametric and additive tree clustering. *Qual. Quant.*, **20**, 169–180.

De Soete, G. (1988) OVWTRE: a program for optimal variable weighting for ultrametric and additive tree fitting. *J. Class.*, **5**, 101–104.

De Soete, G., DeSarbo, W. S. and Carroll, J. D. (1985) Optimal variable weighting for hierarchical clustering: an alternating least-squares algorithm. *J. Class.*, **2**, 173–192.

Fowlkes, B. E., Gnanadesikan, R. and Kettenring, J. R. (1988) Variable selection in clustering. *J. Class.*, **5**, 205–228.

Ghosh, D. and Chinnaiyan, A. M. (2002) Mixture modeling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275–286.

Gnanadesikan, R., Kettenring, J. R. and Tsao, S. L. (1995) Weighting and selection of variables for cluster analysis. *J. Class.*, **12**, 113–136.

Gordon, A. (1999) *Classification*, 2nd edn. London: Chapman and Hall–CRC.

Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.

Hansen, P. and Jaumard, B. (1997) Cluster analysis and mathematical programming. *Math. Program.*, **79**, 191–215.

Hartigan, J. A. (1975) *Clustering Algorithms*. New York: Wiley.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L. and Botstein, D. (2000) Gene shaving: a new class of clustering methods for expression arrays. *Technical Report*. Department of Statistics, Stanford University, Stanford.

Hubert, L. J., Arabie, P. and Meulman, J. J. (2001) *Combinatorial Data Analysis: Optimization by Dynamic Programming*. Philadelphia: Society for Industrial and Applied Mathematics.

Jain, A. and Dubes, R. (1988) *Algorithms for Clustering Data*. Englewood Cliffs: Prentice Hall.

Kaufman, L. and Rousseeuw, P. (1990) *Finding Groups in Data: an Introduction to Cluster Analysis*. New York: Wiley.

Lazzeroni, L. and Owen, A. (2000) Plaid models for gene expression data. *Technical Report*. Department of Statistics, Stanford University, Stanford.

Meulman, J. J., Zeppa, P., Boon, M. E. and Rietveld, W. J. (1992) Prediction of various grades of cervical preneoplasia and neoplasia on plastic embedded cytobrush samples: discriminant analysis with qualitative and quantitative predictors. *Anal. Quant. Cytol. Histol.*, **14**, 60–72.

Milligan, G. W. (1989) A validation study of a variable-weighting algorithm for cluster analysis. *J. Class.*, **6**, 53–71.

Mirkin, B. G. (1996) *Mathematical Classification and Clustering*. Boston: Kluwer.

Späth, H. (1980) *Cluster Analysis Algorithms*. Chicester: Horwood.

Van Buuren, S. and Heiser, W. J. (1989) Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, **54**, 699–706.

Van Os, B. J. (2001) Dynamic programming for partitioning in multivariate data analysis. *PhD Thesis*. Department of Data Theory, Leiden University, Leiden.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001) Model-based clustering and data transformations for gene expression data. In *Proc. 3rd Georgia Tech–Emory Int. Conf. Bioinformatics*. To be published.

Discussion on the paper by Friedman and Meulman

David J. Hand (*Imperial College London*)

Modern statistics generally places emphasis on models at the expense of algorithms, so I particularly welcome this paper. Its appearance in the Society's methodology journal may go some way towards helping to broaden the perception of what statistics is—and what statistics should be. However, there were some aspects of the paper which puzzled me.

In *cluster analysis* we partition a data set with the aim of identifying 'naturally occurring' groups—we seek to 'carve nature at the joints'. Now, I agree that in carving nature at the joints it is unreasonable to expect all the variables which happen to have been measured to contribute to a natural grouping. Presumably, in almost all applications, some of the variables will be relevant and others not. The trick is deciding which variables matter, so that we have a combination of variable selection and traditional cluster analysis. But what are Friedman and Meulman proposing? It is certainly a partition, but it is hardly carving nature at the joints if some points are allocated to cluster A on the basis of their similarity on variable set x while the others are allocated to cluster B on the basis of variable set y. Where are the joints—the gaps which separate cluster A from cluster B?

This lack of joints puzzled me and led me to question the procedure in various ways.

Firstly, by definition, partitioning methods assign each data set object into a unique cluster. But, when the clusters are defined in terms of different sets of variables, there seems to be no reason why we should expect this to be possible: perhaps an object should be in cluster A in terms of one variable set and cluster B in terms of another, without there being a notion of a cluster in terms of the joint set of variables. Should we not, therefore, in principle at least, conduct separate cluster analyses on all subsets of variables? Does not mixing up the subsets simply confuse things?

Secondly, again because it is a partitioning, the authors' method assigns *every* point to a cluster, but it is not at all obvious that this is really what they want to do. In fact, the authors say in Section 11 that

> 'The COSA algorithms attempt to uncover distinct groups of objects that have similar joint values on subsets of the attributes'.

If that is really what the clustering of objects on subsets of attributes (COSA) algorithms did, I would not object: there is nothing here about assigning *all* the objects into groups, but in fact none of the objects in the data set escape from COSA. Let me take this a little further. If some variables are relevant to a cluster for some objects, while many are irrelevant, surely it is entirely reasonable that for some objects none of the measured variables are relevant, i.e. that some of the objects may not lie in any cluster. Why, then, force all objects to lie in some cluster?

This suggests that what the authors are really seeking is not a partitioning tool at all, but rather a tool for detecting locally dense regions of objects and the subsets of variables on which they are locally dense.

A final point: the COSA algorithm 1 lets the weight vector w vary between clusters and the COSA algorithm 2 lets it vary between objects, but another alternative is to let the weight vector vary between pairs of objects, so that w_k is replaced not by w_{kl} or w_{ki} but by $w_k(d_{ijk})$, a function of d_{ijk}. In particular, of course, for a given pair of objects greater weight would be placed on variables for which the two objects were close.

The cluster analysis literature is vast, not least because such problems have been tackled by researchers from a range of intellectual disciplines, including statistics, machine learning, pattern recognition, data mining and others. In fact, over two decades ago, new clustering algorithms were being developed in such quick succession that it was semiseriously proposed that there should be a moratorium on the development of new methods, while we sorted out the properties of the existing ones. This did not happen and development continued, driven by new problems such as dramatic increases in data set size and dimensionality. The key point in developing new methods seems to me to be that they should address some aspect of clustering that has not previously been addressed. Or, if they address old problems, they should do so in a demonstrably superior way. Although I do not necessarily agree with all the points in this paper, the authors are certainly addressing an important issue, and one which is likely to become more important in the future. They have made a substantial contribution to the debate, and it gives me genuine pleasure to propose the vote of thanks.

Chris Glasbey and Dirk Husmeier (*Biomathematics and Statistics Scotland, Edinburgh*)

We welcome this paper, which provides a potentially useful addition to a statistician's toolkit for revealing structure in high dimensional data sets. The approach is algorithmic, and the aim is hypothesis generation rather than validation. A series of heuristic approximations motivates algorithm COSA 2, an algorithm for generating a matrix of distances between all pairs of objects. However, as these distances can be highly non-metric, we would prefer the term 'dissimilarity'.

Although the illustrations in the paper use the distance matrix only with average linkage hierarchical clustering, it could also be input to many other multivariate techniques. For example, non-metric multidimensional scaling would use two- or three-dimensional space, rather than a dendrogram, to visualize clusters of objects. Also, the distance matrix could be input to fuzzy clustering algorithms, to allow for the possibility of overlapping clusters. Targeting is presented as a refinement to clustering of objects on subsets of attributes (COSA) but, as the authors realize, could also be used independently.

One drawback of COSA is that, like classification and regression trees, it is very sensitive to rotation of axes: clusters will be found only if they are parallel to axes. Also, missing values are handled in a less than ideal way in Section 10. Consider a situation where object a has a full set of attributes, object b has a half-set of attributes identical with these and with the rest missing, and similarly object c, but for the other half of the attributes. So it looks like objects a, b and c belong in the same cluster, whereas the distance between b and c is set to ∞, and the dendrogram is forced to be disconnected. If such distances were specified as missing, then this could be handled by most hierarchical clustering algorithms.

The global minimization of objective function (23) is probably NP-hard: hence the adoption of a dual optimization method (Section 5), with a homotopy strategy (Section 6) to avoid becoming trapped in local optima. However, although homotopy requires $\alpha > 0$, the authors appear to set $\alpha = 0$ in the illustrations, with the justification that, when objects cluster only on a very small number of attributes, distance measures (34) and (35) are very similar, as they are both dominated by the relevant attributes. In applications where objects cluster on large sets of attributes, the failure to apply homotopy would have an effect, but then COSA might not be so useful in such cases!

The clustering results can vary substantially as a consequence of changing λ. Its variation can therefore be seen as part of the data mining exercise. However, the problem of finding a 'useful' value for λ may be aided by proposing a probabilistic generative model: an approach which has similarities with Autoclass.

In the illustrations, we did not find it as easy as the authors to identify clusters in the dendrograms. In particular, if the cut points had been chosen slightly differently for the mitochondrial ribonucleic acid (RNA) data, then the match between eight clusters and experiments would have been less than perfect, albeit still very good. So we wonder whether the cut points were at least partly informed by knowledge of which experiment each sample came from. In any case, these experiment-linked clusters may not be of biological relevance to the yeast cell cycle. Natural data sets usually contain alternative and conflicting underlying structures, and it is essential to distinguish relevant structures from irrelevant ones. In the analysis of gene expression data, relevant clusters are usually directly related to a disease under investigation, whereas they are obscured by irrelevant confounding clusters related to normal cellular processes or external experimental conditions. In the case of the mitochondrial RNA data, are the experiments related to different, inherently meaningful biological processes rather than just external conditions, and is the subdivision of some of these experiments into subclusters biologically meaningful?

For the medical data, the authors claim that the clusters they found are related to clinical findings. However, Fig. 11 is not totally convincing, as several of the clusters have very similar distributions of the diagnostic index and we mainly see a partition into only two groups of clusters. This may suggest that most of the clusters found do not actually uncover categories that are relevant for medical diagnostics, but rather alternative, possibly irrelevant, structures.

Finally, we note that, in applications where attributes are commensurable, it may be preferable not to standardize to unit variances. For example, in microarray experiments, genes which show greater variability across samples are often those of greatest interest, so that variability should be preserved and used, such as in 'gene shaving'.

In conclusion, we found this an original and stimulating paper, and it gives us great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

John C. Gower (*The Open University, Milton Keynes*)

In its simplest terms this paper defines distance within each cluster in terms of its own 'metric'. Usually, not only is internal cluster homogeneity stressed but also between-cluster heterogeneity; in this paper I miss the latter. One measure of between-cluster distance defines

$$d_{ij} = \sum_{i \in G_j, j \in G_j} \frac{1}{N_i N_j} D_{ij}(W_K),$$

the average distance between all members of cluster G_i and cluster G_j measured in the metric of cluster G_k, which is especially relevant when $k = i$ or $k = j$. Indeed, d_{ij} and d_{ji} are used in the clustering of objects on subsets of attributes algorithms when updating cluster membership but seem not to enter the primary criteria. For any clustering into L groups we may construct the asymmetric matrix $\mathbf{D}$ of between- and within-group distances. This paper concentrates on the diagonal and tr($\mathbf{D}$) but perhaps ratio criteria using the whole of $\mathbf{D}$ should be considered.

Although I recognize the virtuosity with which the penalty function parameter λ, the homotopy parameter η and the tuning parameter α of the clustering of objects on subsets of attributes algorithms are used, and, although they seem to achieve results, in my opinion they introduce some degree of arbitrariness. Further, common to many cluster problems, we have scaling considerations: attribute weights w_{kl} which sum to 1 in each G_l (is equal weighting of this nature always desirable?), attribute scaling s_k which ignores cluster structure (should it, or should it be updated to give within-cluster normalization?) and the cluster weights W_l first appearing in equation (3). These interact in a complicated way that adds to concerns. These issues are easily raised but less easily resolved. Ratio criteria might help. Scaling effects might be removed by categorizing quantitative variables (see for example Gifi (1990)) and would certainly help with the targeted clustering of Section 11. A full development could view determining attribute weights as an optimal scores problem, probably leading to an intractable clustering problem.

The readiness of clusterers to specify the number of non-hierarchic clusters but not the number of hierarchic clusters puzzles me. Given any l-group criterion C_l, say, it is simple to give an objective criterion for the best clustering into L nested groups as that which minimizes $\Sigma_{l=1}^{L} W_l C_l$ where $G_1, G_2, \ldots, G_L$ are constrained to be nested and the weights W_l are specified. This is a horrendous combinatorial minimization problem but the criterion allows the results of different heuristic algorithms to be compared easily.

Hans C. van Houwelingen (*Leiden University Medical Center*)
First, I congratulate the authors for an interesting and thought-provoking paper. I read it with great pleasure and I am pleased to take part in the discussion. A great part of the paper is devoted to developing search algorithms that lead to cluster solutions minimizing the clustering of objects on subsets of attributes (COSA) criterion. That is not my field of expertise. Therefore, I want to concentrate on the applications of the COSA algorithm and extensions to more structured problems.

Validation of diagnostic subgroups obtained through clustering of objects on subsets of attributes
I appreciate the medical example of Section 12.3. Clusters as obtained here are often used to establish new diagnostic categories. Of course such new diagnostic groups seek validation. External validation as shown in Fig. 11 can be very useful. The classification of new patients into one of the clusters is easily done by a computer but could be difficult to explain to the doctors involved. Therefore, I propose to use supervised learning techniques like discriminant analysis, polytomous logistic regression, classification and regression trees and neural nets to 'predict' the clusters from the available data. This can help to understand the role of the different attributes, to 'soften' the clusters and to answer the questions about sensitivity, specificity and predictive value that can be expected from the medical field.

Extension to sequential or longitudinal data
The biological interest in gene expression data is often in the clustering of genes, using expression under different conditions. There are no such examples in this paper. Is it feasible in COSA to cluster 10 000 objects on 20 attributes, say? The different conditions that are studied in microarray experiments are often of a sequential nature: samples obtained at different time points in an experiment. Such data are very similar to longitudinal data. I have always been impressed by the complexity of growth curve data and Theo Gasser's pioneering work in understanding them (Gasser and Kneip, 1995). I would like to see COSA applied to such data. In COSA no relationship is assumed between the different attributes. To make it more fit for longitudinal data, it would be useful to introduce a concept of smoothness into COSA. I wonder whether it would be feasible to introduce a smoothness penalty in the algorithm that would lead to clusters based on sets of neighbouring attributes.

J. B. Bugrien (*University of Leeds*)
I thank the authors for an interesting paper that makes a valuable contribution to cluster analysis techniques. However, focusing on subsets of variables seems to ignore the effects of correlation between the variables.

Here is a simple example where the proposed 'clustering of objects on subsets of attributes' method does not work. Consider two clusters in two dimensions (with co-ordinates labelled x and y) lying on two closely spaced parallel lines at a 45° angle to the x-axis. For example suppose that points in the first cluster take the form $(Z_i, Z_i + \varepsilon), i = 1, \ldots, N$, and points in the second cluster take the form $(Z_i, Z_i - \varepsilon), i = N + 1, \ldots, 2N$, where Z_i are independent and identically distributed uniform variables on $[-1, 1]$ and $\varepsilon > 0$ is small. Recall that for $U, V \sim \text{Unif}[a, b]$ we have $E|U - V| = (b - a)/3$. Then the scale measure reduces to $s_k \simeq \frac{2}{3}, k = 1, 2$. With equal weights $w_1 = w_2 = \frac{1}{2}$, the clustering criterion $Q(\cdot)$ from equation (3) for the true clustering c reduces to $Q(c) \simeq 2$.

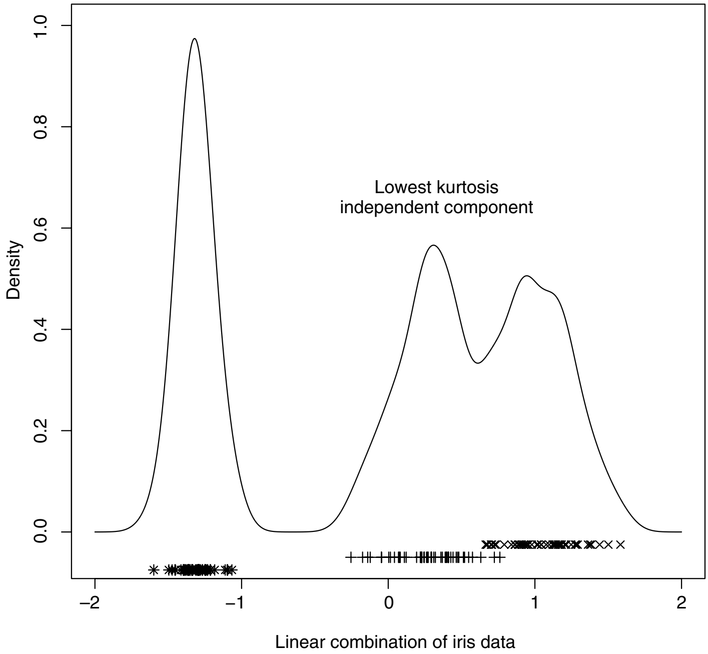

Fig. 12. Plot of the density estimate for the linear combination of minimum kurtosis from the iris data: $*$, *setosa*; +, *versicolor*; ×, *virginica*

However, if we look at a false clustering c^* defined by two clusters $x \geqslant 0$ and $x < 0$, we obtain scale measure $s_k \simeq \frac{1}{3}, k = 1, 2$, and a much smaller clustering criterion value $Q(c^*) \simeq 1$. Thus, in this case the clustering criterion $Q(\cdot)$ cannot pick out the true clustering c.

Another approach to clustering which takes correlation into account is given by 'independent component analysis' (e.g. Hyvarinen *et al.* (2001)), which is closely related to projection pursuit. The objective is to look for linear combinations of the data which are as non-Gaussian as possible. In one version, we look for large values of the absolute kurtosis $|\kappa|$, with $\kappa = 0$ corresponding to Gaussianity. In fact, though it has not been widely recognized, clustering is usually associated with sub-Gaussianity, $\kappa < 0$, rather than, super-Gaussianity, $\kappa > 0$.

To illustrate the use of kurtosis, consider the classic iris data that were analysed by Fisher (1936) with $n = 150$ observations in $p = 4$ dimensions on three species (*setosa*, *versicolor* and *virginica*). It is well known that iris *setosa* is well separated from *versicolor* and *virginica*, which are less well separated from one another. If we ignore the group membership, independent component analysis can be used to find the linear combination with minimum kurtosis. A kernel density estimate of the data on the resulting linear combination is shown in Fig. 12. As expected the plot clearly separates *setosa* from *versicolor* and *virginica*. Somewhat fortuitously, it also nearly separates *versicolor* from *virginica* as well. In fact, it is very close to the first canonical variate from traditional discriminant analysis, which is constructed using the group membership.

Guy Nason (*University of Bristol*)
I congratulate the authors for the fascinating and stimulating paper that we have heard tonight.

One of my pet projects is the statistical analysis and interpretation of early day motions (EDMs) data (see Nason (2001) or Berrington (1973)). EDMs are small manifestos that can be introduced by any UK Member of Parliament (MP), who then accumulates signatures from other MPs who agree. EDMs are not whipped. Party machines do not control who does or does not sign. Their unwhipped nature is a key argument for proposing that EDMs provide a more reliable source of information about MPs' intentions than, say, their voting record or public pronouncements. As an experiment we applied clustering of objects on subsets of attributes (COSA) to incidence data from the 1998–1999 session of Parliament which indicates by 1 or 0 whether or not any of 549 MPs signed any one of 1169 EDMs.

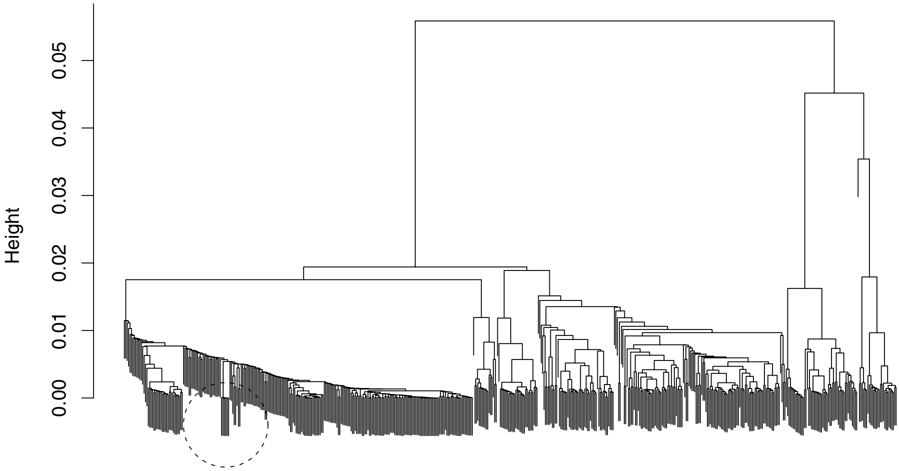

Fig. 13. Dendrogram after applying default hierarchical clustering to output from COSA applied to the EDM data

After applying COSA and the default hierarchical cluster analysis we obtained the dendrogram as shown in Fig. 13. There are many interesting groupings to be explored in Fig. 13 but for the moment consider the group of MPs that is indicated by the dotted circle which contains the special Conservative MPs James Arbuthnot, Damian Green, William Hague, Peter Lilley, Theresa May and David Willetts. During 1998 William Hague was the party leader and Arburthnot was his Chief Whip. Lilley, Willetts, May and Green were all part of Hague's shadow cabinet at some time during 1998 (indeed the last three were Secretary of State and front bench spokesmen for Education and Employment).

Clearly this group seems to be a real grouping but COSA can go further and ask what EDMs make this grouping? COSA reports that 57 EDMs have importance of 20 and all else has zero importance. None of the Hague group signed any of the 57 EDMs although they were not the only EDMs that they all did not sign. One has a flavour of the type of these 57 EDMs by looking through their subject headings, e.g. EDM 5 'Hunting wild mammals' (anti-hunting), EDM 11, 'Wildlife charter and wildlife Bill' (wildlife protection), and EDM 15, 'Freedom to roam' (enabling people to roam over parts of the UK that they were not previously able to). These three EDMs, at least, attracted widespread support with 192, 349 and 214 signatories. This level of support is unusual for most EDMs.

Of course, more work needs to be done both on this application and on understanding of COSA. However, it is clear already that COSA is a valuable and important contribution to statistics and that tonight's authors should be thanked.

The following contributions were received in writing after the meeting.

Frank Critchley (*The Open University, Milton Keynes*)
In welcoming the authors' highly innovative methodology, developed to address pressing practical problems, I offer the following remarks and suggestions, regretting that I am unable to do so in person. These focus on two topics: *goals* and *evaluation*.

Goals

The paper conceives of clusters solely as groups of objects having *internal cohesion*. The twin property of *external isolation*, which is used in much of the literature, is ignored. This leads, in particular, to a somewhat indirect strategy for defining distances between all pairs (Section 5).

A natural suggestion is to work throughout with loss functions Q defined as sums over *all* pairs (l, m), small values of $l \neq m$ terms reflecting large separation between different groups.

Evaluation

The methodology that is introduced in the paper requires a number of quantities to be specified, including the distances in equation (5), the attribute–plurality incentive function (19) and its key parameter λ, the majorizing distance (30), the (hierarchical) clustering algorithm, the homotopy parameter α, the number of nearest neighbours K and the attribute importance parameter ε. Especially in view of the complexity of the function whose minimum is sought, different combinations of these specifications have the potential to produce very different results. This being the case, evaluation of the validity and usefulness of the output is especially important. In their closing paragraph, the authors assert that this can *only* be done separately in each application context. This seems both to run familiar risks of circularity, especially in exploratory contexts, where little relevant background knowledge is available (possible instances occurring in the discussion of the choice of α and K), and to be unnecessarily restrictive—diagnostic, cross-validation and related technologies being adaptable for use here.

Indeed, a further *substability* check suggests itself here. The basic idea is very simple: confidence in the validity of a subset of groups, and of the attribute subsets on which they form, is increased the more stable they are under reanalysis following deletion of other attributes having small importance for them and/or of other groups of objects.

Peter D. Hoff (*University of Washington, Seattle*)

My comments concern the types of clusters that clustering of objects on subsets of attributes (COSA) algorithm 2 identifies, and a model-based alternative. The simulated data that were presented in the paper confound a mean shift of 1.5 with a fivefold reduction in standard deviation. To separate these effects I ran COSA algorithm 2 on two simulated data sets which were similar to those in the paper. Both consisted of 10000 attributes on 100 objects, all attributes being standard normal except for the first 150 attributes of a 15-object group. In the first data set these attributes were normally distributed with mean 1.5 and standard deviation 1, and in the second with mean 0 and standard deviation 0.2. The R code that was used to generate these data is available at `www.stat.washington.edu/hoff/public/COSA/cosa.example.r`. Dendrograms based on COSA algorithm 2 distances are shown in Fig. 14. The algorithm picks up the change in variance but not the mean. This suggests that the clustering results in the paper are predominantly due to the objects in group 2 having measurements at 150 attributes that are all tightly concentrated around common values, and less a result of these common values differing from those in the other group.

Suppose that we *are* interested in identifying clusters based on shifts in the means of subsets of attributes, as described by the model

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\gamma}_{(c(i))} + \boldsymbol{\varepsilon}_i,$$

$$\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \overset{\text{IID}}{\sim} \text{multivariate normal}\{\mathbf{0}, \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)\},$$

where the N-dimensional 'mean shifts' $\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(L)}$ represent the differences in means between L groups and a base-line mean $\boldsymbol{\mu}$. To allow for clustering on subsets of attributes, we model $\boldsymbol{\gamma}_{(l)} = \mathbf{s}_{(l)} \times \boldsymbol{\delta}_{(l)}$, where $\mathbf{s}_{(l)}$ is a binary sequence, $\boldsymbol{\delta}_{(l)}$ a vector of real numbers and '×' represents elementwise multiplication.

Estimation for such a model is reasonably straightforward by using a Dirichlet process mixture model which extends the work of MacEachern (1994): uncertainty about $\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(L)}$ and the clustering can be represented simultaneously by a Dirichlet process prior for the values of $\boldsymbol{\gamma}_{(l)} = \mathbf{s}_{(l)} \times \boldsymbol{\delta}_{(l)}$. Such an approach can provide an estimated clustering and a measure of uncertainty, and it can be readily adapted to model other data features such as correlated attributes, heteroscedasticity or t-errors. Such an approach is

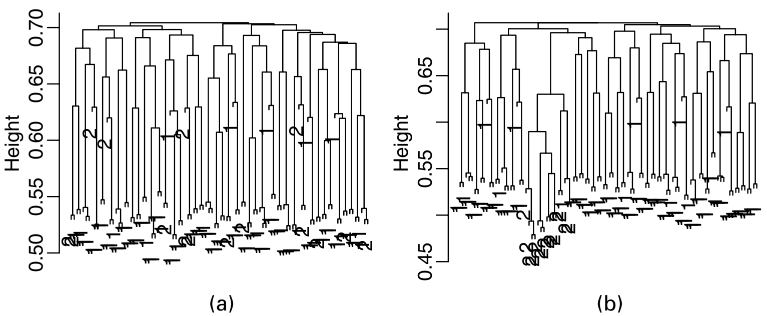

Fig. 14. Average linkage dendrograms based on the untargeted COSA algorithm 2 distances: the data were generated from a standard normal distribution except for 150 attributes of 15 objects, having (a) mean 1.5 and standard deviation 1 and (b) mean 0 and standard deviation 0.2

outlined in Hoff (2004), where it is seen to outperform other clustering methods for these types of data sets, and is extended to model spatially correlated genomic data.

G. J. McLachlan and R. W. Bean (*University of Queensland, Brisbane*)
The authors are to be congratulated on tackling the challenging problem of clustering very high dimensional data. Their approach to find subgroups of objects that cluster on subsets of the attributes (variables) rather than on all of them simultaneously is relevant in many practical situations, such as in the example that they give on the clustering of a limited number of tissue samples on the expression levels of thousands of genes.

The approach is similar to that adopted by McLachlan *et al.* (2002) in considering the latter problem via their EMMIX-GENE procedure. To cluster the tissue samples, McLachlan *et al.* (2002) clustered the variables into groups, on the basis of Euclidean distance, so that highly correlated variables tend to be put into the same group. Before this step of grouping the variables, the highly correlated variables are standardized and there is also the provision of eliminating variables that are considered to be of little use (individually) in clustering the data (in terms of fitted mixture models). Having grouped the variables into, say, n_0 groups, the EMMIX-GENE program displays heat maps of the expression levels for the tissues for the genes within a group, which provides a visual aid in assessing which groups of genes will lead to similar clusterings of the tissues. This question can be addressed more formally by fitting mixture models on the basis of the genes within a group using, if necessary, mixtures of factor analysers to handle the problem of a large number of variables relative to a limited number of objects. Alternatively, one can proceed by working with a representative of each group such as its sample mean (a metagene). We applied the EMMIX-GENE procedure to the mitochondrial ribonucleic acid data set. A reduced number of genes (4978) was clustered into $n_0=40$ groups, and then the tissues clustered on the basis of the top 10 metagenes. Seven of the implied clusters for a nine-component mixture model corresponded to the experiments designated as Hol, Cho, Spe_alpha, Spe_elut, Spe_cdc, Chu and Der, whereas the other two corresponded to Vel and to Myers and Roth combined.

Finally, in their discussion of non-distance-based modelling methods, the authors provide some references in the 1990s on product density mixture modelling, which they note is the closest in spirit to their approach. Some earlier references on the mixture modelling approach to clustering are Wolfe (1967), Day (1969), Ganesalingam and McLachlan (1979) and McLachlan and Basford (1988).

The **authors** replied later, in writing, as follows.

We thank the discussants for their valuable comments and important suggestions. They provide further insight into the nature of the clustering of objects on subsets of attributes (COSA) procedure and its results. Within space limitations, we respond to some of the issues raised.

Hand questions the nature of the groupings to which COSA is sensitive, suggesting that it may not be sensible to define objects as being in separate groups solely because they tend to concentrate on different variable subsets. He would add the additional requirement that the groups be separated by regions of low object density in the space of joint attribute values ('joints' or 'gaps'). COSA is also sensitive to this latter type of clustering. Whether the former type of grouping represents distinct phenomena of interest depends on the context of the problem as interpreted by the domain expert, who may want at least to be aware that the data exhibit this type of structure. Inspection of the relationships between the uncovered groups can reveal the extent to which there are or are not such gaps between them and, if not, whether the groupings have domain relevance. It should be noted that the goal of mixture modelling is also to uncover both types of groupings without particular reference to gaps, whenever the respective component covariance matrices are not constrained to be identical.

Hand also questions whether it makes sense to apply partitioning, where all objects are assigned to unique groups, in the presence of the former type of grouping. As noted by Glasbey and Husmeier, COSA can be used with other distance-based multivariate procedures as well, such as fuzzy clustering or multidimensional scaling. Moreover, using hierarchical clustering methods or multidimensional scaling we can identify clustered subgroups whose union need not cover the entire data set. For example, from Figs 2(c), 2(f) and 2(i) we might infer that there is a single small group of 15 objects; the remaining 85 objects represent a diffuse unclustered background. For the data set that is represented in Fig. 4 we might conclude that there are no clustered groups, only an unstructured background. Here those inferences in fact characterize the models from which the respective data sets were generated. Thus we need not assign every point to a cluster, and we would still maintain that 'the COSA algorithms attempt to uncover distinct groups of objects that have similar joint values on subsets of the attributes', an objective to which Hand does not object.

We favour the idea of using multidimensional scaling (MDS), a spatial representation, *in addition* to hierarchical clustering since it reveals the tightness within clusters and the distances between different clusters in a straightforward way. In this manner, MDS provides useful information when different clustering methods do not provide exactly the same clusters. Also, MDS gives a very nice spatial representation of the influence of the homotopy parameter η for $\alpha > 0$. Some modest experimentation showed that the overall cluster structure often does not change but that clusters do seem to become tighter for increasing values of η, which would hopefully answer the concern of Glasbey and Husmeier with respect to the use of $\alpha = 0$, resulting in $\eta = \lambda$ in the paper. Larger values of η that are obtained by setting α to a small value are helpful when the cluster structure is less clear at the outset.

At the time, we did consider Hand's suggestion that different weightings be applied to pairs of objects, rather than just to individual objects from which the paired distances are derived (equations (30) or (33)). However, we could not derive a computationally feasible search strategy for this approach, especially for large high dimensional data sets that characterize many of the current applications of clustering.

Both Bugrien, and Glasbey and Husmeier point out that COSA has reduced sensitivity to clusters that primarily concentrate on linear combinations of the attributes, as opposed to the actual attributes themselves. This is a clear design limitation dictated by the difficulties that are associated with large high dimensional data sets. Bugrien suggests projection pursuit or independent components analysis as being sensitive to linear combination clustering. Gaussian mixture modelling with general covariance structures is another possibility. Indeed, for smaller, low dimensional data sets these are viable alternatives that can often be successful, as nicely illustrated by Bugrien on the iris data. For modern high dimensional data, such as produced in genomics and proteomics applications, these approaches run into both estimation and computation difficulties as discussed in Section 13. Furthermore, interpretability is an important aspect of any exploratory data analysis tool, and linear combinations of variables are much more difficult to interpret than the original variables, especially those involving many variables. We believe that this lack of interpretability is one of the principal reasons why projection pursuit (and classical canonical correlation analysis) have failed to become widely used tools for exploratory data analysis.

Both Gower and Critchley observe that the COSA procedure involves several parameters whose values can affect the results. As Critchley, and Glasbey and Husmeier remark, the most important parameter is the scale or smoothing parameter λ that directly enters the motivating criterion (23). The others are associated with the search strategy. Most multivariate procedures involve some type of smoothing parameter and, as noted by Glasbey and Husmeier, its variation can be viewed as part of the data mining exercise. We note that other techniques intended for high dimensional data sets involve dimension reduction screening as a preprocess necessary to reduce the number of variables that is input to the clustering algorithm, as for example in the procedure described by McLachlan and Bean. This preprocessing involves many explicit as well as implicit parameters to be specified that affect the results of the overall procedure.

We agree with Critchley that it is important to assess stability with respect to the search strategy parameters (homotopy parameter α and number of nearest neighbours K in COSA algorithm 2). However, we do not quite understand the substability check. It is reasonable to expect that results be stable under deletion of variables that have small importance for all the uncovered groups. Removing variables that are relevant to the clustering of some groups but not to others would be expected to change the results.

Gower makes some interesting and important observations concerning criteria (16) and (23) that motivate the COSA procedures. Along with Critchley, he notes that our criterion stresses within-group homogeneity and suggests alternative criteria that involve between-group heterogeneity as well. We agree that such alternatives might have the potential to improve performance. The difficult part is developing feasible search strategies for their optimization. It may turn out to be possible to do so, providing interesting avenues for future research.

Gower also notes that our choices for scaling the variables and normalizing the weights are not the only possibilities. Although normalizing the variable weights that are associated with each group to sum to 1 seems natural, this is by no means the only alternative. Different choices will give rise to different criteria for which it may be possible to develop feasible search strategies.

Hoff notes that COSA is sensitive to differences in scale as well as location among the clustered groups and suggests that it is sensitivity to the former that predominately drives the method. He proposes a model-based procedure that is intended to be sensitive only to location differences, the implication being that such a procedure should have increased power when the clustered groups differ primarily in location. This may be so in some situations but we miss this being demonstrated in the example provided. We are not as optimistic concerning the straightforward nature of the estimation procedure when computational feasibility is considered. It remains to be seen whether fast search algorithms can be developed for this type of model that will allow application to large high dimensional data sets.

McLachlan and Bean describe their EMMIX-GENE procedure, which they regard as being similar to COSA. Although their goals may be similar, and sometimes they can produce similar results, we feel that there are substantial characteristic differences between the two approaches. As with all mixture modelling, EMMIX-GENE first preprocesses the variables to reduce their number so that model estimation is feasible. This is done in two stages; first a subset of the variables is selected on the basis of individual variable marginal distributions only. This presumes that the high dimensional clustering will be detectable in the marginal distributions of all the variables that contribute to the clustering structure. This can but need not happen, as illustrated Fig. 1. Next the variables are clustered on the basis of their correlations. This presumes that the clustering of the objects will be sufficiently strongly reflected in the (linear) correlational structure of the variables. Again this often, but not always, happens. For example, in the data that are described in Section 12.1, the variables all have very weak (population) correlation, and for the data depicted in Hoff's Fig. 14(b) they are completely uncorrelated. In contrast, COSA can be applied to the entire data set without any preprocessing, thereby remaining sensitive to clustering that is not necessarily reflected in the marginal distributions or correlational structure of the variables.

We thank McLachlan and Bean for providing the lacking references to the early literature on mixture modelling.

Glasbey and Husmeier are not convinced that the clusters found in the medical example refer to categories that are relevant to medical diagnosis. We can answer here that the unsupervised COSA clustering does slightly better than a supervised classification tree analysis for differentiating the five given diagnostic categories. Moreover, as van Houwelingen remarks, clusters as obtained in the medical example are often used to establish *new* diagnostic categories. He recommends using supervised learning methods to 'predict' the clusters uncovered. This is a very good idea since it provides additional information, along with COSA's variable importance measure, concerning the role of the attributes in differentiating the groups that might be instrumental in answering questions from the medical field.

van Houwelingen also asks whether COSA could be used for straightforward clustering of the genes using expression under different conditions. Indeed, although the computational requirement is somewhat higher, it is feasible to use COSA to cluster, say, 10000 objects on 20 attributes rather than 20 objects on 10000 attributes. In this application of COSA, the goal would be to uncover groups of genes that cluster on subsets of samples, rather than groups of samples clustering on subsets of genes. Although these goals have some similarity they are not identical and additional useful information may be obtained from the former exercise.

van Houwelingen's suggestion of introducing smoothness as a constraint in the context of clustering is very interesting and potentially quite powerful. This can be accomplished in COSA by using an approach originated by Friedman and Popescu (2004) in the context of supervised learning. With COSA, smoothness on the attributes can be achieved by constraining the attribute weights (w_{kl} equation (22) or w_{ki} equation (43)) to be smooth functions of the variable index k. This can be accomplished by replacing S_{kl} or S_{ki} (equations (12) and (24) respectively) by corresponding smoothed quantities; for example

$$\{\tilde{S}_{ki}\}_{k=1}^{n} = \text{smooth}_{\gamma}(\{S_{ki}\}_{k=1}^{n})$$

would replace $\{S_{ki}\}_{k=1}^{n}$ in equation (43). Here $\text{smooth}_{\gamma}(\cdot)$ represents an operator that outputs a smoothed version of an ordered sequence of input values and γ is a parameter that regulates the degree of smoothness imposed. Any convenient univariate smoother could be employed. Similarly, we could impose a smoothness constraint on the *objects* in equation (43) by smoothing with respect to the index i. This might be appropriate if the objects represented samples taken sequentially in time. One could even *jointly* smooth with respect to both indices using a bivariate smoother, if appropriate, We intend to investigate these approaches and, if they prove successful, to implement them as options in future versions of the COSA software. The same applies to a different treatment of missing data and alternative scaling of the attributes that various discussants suggested.

Finally, we appreciate Nason's sharing his experience with COSA as applied to the early day motions Parliament data. We are glad that he found COSA to be useful and it is our hope that others will as well.

References in the discussion

Berrington, H. (1973) *Backbench Opinion in the House of Commons 1945–55*. Oxford: Pergamon.

Day, N. E. (1969) Estimating the components of a mixture of two normal distributions. *Biometrika*, **56**, 463–474.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

Friedman, J. H. and Popescu, B. E. (2004) Gradient directed regularization for linear regression and classification. *Technical Report*. Department of Statistics, Stanford University, Stanford.

Ganesalingam, S. and McLachlan, G. J. (1979) A case study of two clustering methods based on maximum likelihood. *Statist. Neerland.*, **33**, 81–90.

Gasser, T. and Kneip, A. (1995) Searching for structure in curve samples. *J. Am. Statist. Ass.*, **90**, 1179–1188.

Gifi, A. (1990) *Nonlinear Multivariate Analysis*. Chichester: Wiley.

Hoff, P. D. (2004) Clustering based on Dirichet mixtures of attribute ensembles. *Technical Report 448*. University of Washington, Seattle.

Hyvarinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis*. New York: Wiley.

MacEachern, S. N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communs Statist. Simuln Computn*, **23**, 727–741.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering*. New York: Dekker.

McLachlan, G. J., Bean, R. W. and Peel, D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Nason, G. P. (2001) Early Day Motions: exploring backbench opinion during 1997–2000. *Technical Report 01:11*. Statistics Group, Department of Mathematics, University of Bristol, Bristol.

Wolfe, J. H. (1967) NORMIX: computational methods for estimating the parameters of multivariate normal mixtures of distributions. *Research Memorandum SRM 68–2*. US Naval Personnel Research Activity, San Diego.