Mixture Model Clustering using the *Multimix* Program

Lynette Hunt and Murray Jorgensen
University of Waikato
Hamilton, New Zealand

Abstract

Hunt [1996] has implemented the finite mixture model approach to clustering in a program called *Multimix*. The program is designed to cluster multivariate data with categorical and continuous variables and possibly containing missing values. In this paper we describe the approach taken to the design of *Multimix* and how some of the statistical problems were dealt with. As an example of the use of the program we cluster a large medical dataset.

Key Words: Cluster analysis, EM algorithm, Latent class analysis, Local independence, Multivariate Normal distribution, Location model, Prostate cancer data.

1 Introduction

This paper is concerned with the statistical analysis of multivariate data from a mixture of finitely many populations when there is no information about membership in any component population. This is known as *cluster analysis* or *unsupervised learning*. The goal is to partition the sample into groups so that members of a group are as similar as possible. This is usually done by any one of a number of deterministic algorithms the most common of which we discuss below.

Cluster analysis is different from *Discriminant Analysis* where it is possible to classify members of a random sample from a mixture of populations according to which population they come from. See Mardia et al. [1979, Chapter 11] and McLachlan [1992] .

There are many different methods for cluster analysis. These methods can be broadly categorised as hierarchical or non-hierarchical. Clustering using hierarchical methods, is generally obtained through either agglomerative algorithms, which begin with a cluster for every observation and successively merge clusters, or divisive algorithms which begin with a single cluster and which continually split clusters.

It is possible to visualise two extremes, one in which each object is considered to be a single member cluster, and one in which all n objects are contained in a single cluster. Each cluster obtained at any stage in the procedure is a combination or division of clusters at other stages. A hierarchical strategy finds an efficient path between these two extremes.
Once an object is assigned to a cluster under a hierarchical strategy, there is no provision for reallocation of the objects that have been poorly allocated at an earlier stage in the process. Each stage of the analysis involves the computation of the cluster similarity (or distance) matrix. Since the clusters at any stage are obtained by the fusion (agglomerative methods), or division (divisive methods) of clusters from the previous stage, these methods lead to a hierarchical structure of the objects. This is represented by a dendogram, also known as a tree diagram.

1.1 Similarity matrix clustering techniques

Hierarchical clustering techniques are usually implemented with the data represented by a matrix of proximities (d_{ij}), where d_{ij} is the proximity of observations i and j. The proximity d_{ij}, can either be a similarity or a dissimilarity measure. To convert a dissimilarity into a similarity index we may, for example, divide it by the greatest dissimilarity observed in the data and subtract this from 1.
Proximities may be obtained in various ways, one method being to ask a number of people to subjectively assess all pairs of observation in a small set for degree of similarity, recording the answer as a number between 0 (least similar) and 1 (most similar). The similarities for analysis can then be obtained by averaging the subjective similarities over the panel of judges.
More often each observation has a number of measured attributes or variables, often at differing levels of measurement (binary, nominal, ordinal, interval or ratio), and we require some means of calculating proximities from the data. Anderberg [1973, Chapters 4 & 5] and Gordon [1981, Chapter 2] give surveys of many methods of calculating proximities for the case of a single variable. Typical examples are the euclidean distance for interval variables and the Jacard coefficient $n_{11}/(n_{11}+n_{10}+n_{01})$ for binary (0,1) data. Another binary coefficient is the simple matching coefficient $(n_{11}+n_{00})/(n_{11}+n_{10}+n_{01}+n_{00})$; indeed Anderberg lists 14 possibilities, though deprecating 5 of these. Which notion of proximity makes the most sense depends on subject area considerations.
Once a similarity measure s_{ijk} comparing observation i with observation j has been selected for each attribute k they may be combined, essentially by averaging over the attributes. In the case of a rare binary attribute k we may wish to exclude s_{ijk} from the average as uninformative about the similarity of i and j. Details about combining similarity measures in this way are given by

Gower [1971].
Agglomerative hierarchical techniques differ primarily in how they measure the distance or similarity of two clusters, where a cluster may at times, consist of a single observation only. For example, the Euclidean distance d_{ij} between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as $d_{ij} = [(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)]^{1/2}$, while the Mahalanobis distance is defined as $d_{ij} = [(\mathbf{x}_i - \mathbf{x}_j)'\hat{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)]^{1/2}$, where $\hat{\Sigma}$ is the within cluster covariance matrix. Further details on the properties of these distances and other distance measures are given by Mardia et al. [1979] and Gordon [1981].
In single linkage (nearest neighbour) clustering, the distance between two clusters is defined as the distance between their two nearest neighbours:

$$d_{AB} = \min_{\substack{i \in A \\ j \in B}}(d_{ij})$$

where d_{AB} is the dissimilarity between two clusters A and B and d_{ij} is the dissimilarity between two observations i and j. This technique can lead to 'rod' type elongated clusters.
With complete linkage (farthest neighbour) clustering, the distance between two clusters is defined as the distance between their two furthest neighbours:

$$d_{AB} = \max_{\substack{i \in A \\ j \in B}}(d_{ij}).$$

This method tends to produce compact clusters.
Other standard linkage methods replace the "min" and the "max" of the previous methods by measures of central tendency.
Lance and Williams [1967] give a general agglomerative algorithm with which many of the common hierarchical linkage methods can be described. If two groups r and s amalgamate to form a new group t, the dissimilarity between this group and any other group can be expressed in an equation form. Gordon [1981] includes their table of the algorithm parameters for different techniques.
With hierarchical clustering, the number of clusters is obtained by selecting one of the clusterings in the nested sequence of groupings displayed in the dendogram. The most common method used is to examine the dendogram for large changes in the distance or dissimilarity between adjacent fusion levels. A "large" change when going from K to $K-1$ groups might be indicative of K groups. This criterion is somewhat subjective. Other sources of subjectivity lie in the choice of similarity metric for each attribute and the choice of linkage method.
These methods are widely implemented in statistical packages and can be useful for preliminary exploration of small multivariate datasets, especially in combination with visualization techniques such as a plot of the first two principal component scores and a Minimal Spanning Tree [Gower and Ross, 1969].

They are less satisfactory with large data sets (hundreds rather than tens) because of the large number of pairwise similarities which must be processed, and because of the enhanced possibilities for unfortunate and irreversible amalgamations of clusters at an early stage.
An important problem with the use of these forms of cluster analysis lies in the many ways in which the subjective decisions made by the analyst may influence the outcome. The analyst must choose

1. the form of the proximity index
2. the linkage method, and
3. the similarity level at which to 'cut' the dendrogram, or equivalently, the number of groups.

1.2 Optimisation based techniques

Nonhierarchical techniques of cluster analysis have the same extremes as hierarchical techniques, that is, n clusters consisting of one observation and one cluster with all n observations in it. However nonhierarchical techniques allow points to be reallocated to other clusters during the clustering process. These techniques of cluster analysis often use optimisation procedures in which observations are transferred between clusters with the aim of optimising some clustering criterion that rewards both within-cluster similarity and between-cluster differences. Once again, there are many different methods available because of different optimising criteria and different optimising algorithms. For further discussions on these procedures see for Everitt [1980] and Hand [1981].
The k-means algorithm of Hartigan [1975] is a commonly used optimisation technique. The means of each of the k initial clusters are found, and then each data point is examined to see if it is closer to the mean of another cluster than to the mean of its current cluster. If this occurs, that point is transferred and the cluster means are recalculated. The means can be recalculated after each data point has been reallocated, or after all the data points have been examined and those that needed reallocating have been transferred. The means of the k clusters are calculated and the process is repeated. In this procedure, the cluster mean is the point that minimises the sum of squares of the distances (to that point) of the observations in that cluster.
The "classification likelihood" approach is a nonhierarchical technique that uses a form of likelihood function as a clustering criterion. Under this approach, a probabilistic formulation is taken in which it is assumed that the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ each arise from any one of K possible sub-populations with a probability density function of $f(\mathbf{x}; \boldsymbol{\theta}_k)$ for $k = 1, \ldots, K$. This approach differs from the discriminant analysis problem in that it is not known which

sub-population the observation comes from. Let

$$z_{ik} = \begin{cases} 1 & \text{if observation } i \in \text{group } k; \\ 0 & \text{if observation } i \notin \text{group } k, \end{cases}$$

and define the vector of indicator variables as $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$. The likelihood function is given by

$$L_{Class}(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{f(\mathbf{x}_i; \boldsymbol{\theta}_k)\}^{z_{ik}}$$

Let $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and $\boldsymbol{\phi} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. Maximisation of $L_{Class}(\mathbf{z}, \boldsymbol{\phi})$, the log-likelihood for the complete data is with respect to $\boldsymbol{\phi}$ and $\mathbf{z}$. That is, the unobservable indicator variables $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are treated as unknown parameters to be estimated along with $\boldsymbol{\phi}$. The maximisation process can be carried out by computing the maximum value of the likelihood over all possible partitions of the n observations to the K groups. This approach was considered by several authors including Scott and Symons [1971], Sclove [1977] and Symons [1981]. More recently Banfield and Raftery [1993] have extended the methods of Scott and Symons [1971] and their approach is discussed below. Unfortunately with this procedure, the z_{ij} increase in number with the number of observations, and the maximum likelihood estimates are not consistent [McLachlan and Basford, 1988].
Using the classification likelihood approach, Scott and Symons [1971] showed that the assumption that $\mathbf{x}_i \sim N(\boldsymbol{\mu}_k, \Sigma)$ for $k = 1, \ldots, K$, led to the cluster analysis procedure based on minimising $\mid W \mid$, the determinant of the pooled within group dispersion matrix. This method of cluster analysis was discussed by Friedman and Rubin [1967]. Scott and Symons [1971] found that this approach has the tendency to divide the data into clusters of equal size if the separation between the sub populations is not large. Marriot [1975] pointed out that the maximum (classification) likelihood estimates are not consistent under the assumption of underlying normal distributions with a common covariance structure. Bryant and Williamson [1978] showed that the approach can also be expected to give biased results. Symons [1981] and Binder [1978] give Bayesian versions of this method.
Although usually considered as nonhierarchical clustering techniques, criterion optimization methods may be used in a hierarchical fashion by applying the algorithm repeatedly to subdivide clusters found earlier. Such an approach usually leads to clusters that are not themselves optimal on the criterion.

1.3 Clustering methods based on finite mixture models

There is a vast quantity of literature available on algorithmic cluster analysis. For comprehensive reviews of clustering techniques see Cormack [1971],

Everitt [1980], Jardine and Sibson [1971], and Gordon [1981]. For clustering algorithms see Hartigan [1975] and James [1985].
There are some inescapable drawbacks shared by all these traditional approaches to clustering: any randomness in the sample is not reflected and small perturbations in the sample may lead to quite different groups being formed. Further, experience with real mixed populations shows that they are quite often substantially overlapping, whereas by design most traditional clustering algorithms will tend to come up with compact nonoverlapping clusters.
An alternative to algorithmic cluster analysis, is to adopt a statistical formulation similar to that of discriminant analysis, and regard the observations to be clustered as a random sample from a finite mixture of distributions. However, unlike discriminant analysis, the observations are not identified as belonging to a particular group, and there is often very little information about the form of the population distributions for each group. By making generic distributional assumptions we obtain a well specified model, whose parameters can be estimated by the method of maximum likelihood. The estimated conditional probabilities of group membership can be estimated by Bayes rule using the parameter estimates. These probabilities can be used when the algorithm has converged to obtain a probabilistic assignment of observations to clusters.
Furthermore, the estimated component distributions together with the estimated proportions for each component provide a concise description of what may be a very complicated set of data.
As with any clustering method, clustering by finite mixture models also imposes a structure on the data. It is possible to check the overall fit of the mixture model to the data, although the individual components cannot be checked unless the groups turn out to be well separated. The mixture likelihood approach can be seen as an example of a nonhierarchical clustering technique. But a unification with the mainstream of statistical modelling is achieved because clustering methods based on mixture models allow estimation and hypothesis testing within the framework of standard statistical theory [Aitkin et al., 1981].

2 Earlier work in mixture model clustering

The *Multimix* program to be described later in this paper builds on earlier approaches and is most easily understood as an extension and unification of some of these. The estimation problem for finite mixtures of normal distributions has quite a lengthy history. We will describe some of this work now.

2.1 Mixtures of normal distributions

Karl Pearson put forward a solution in the case of a mixture of two univariate distributions with unequal variances using the method of moments [Pearson, 1894]. This was a difficult problem and involved the solution of a ninth degree polynomial equation. Later investigation showed that likelihood estimation was more efficient than the method of moments for this problem [Tan and Chan, 1972].

Maximum of likelihood estimation for the parameters in mixture distributions was suggested by Rao [1948] , who used Fisher's method of scoring for the estimation of parameters in a mixture of two univariate normal distributions with equal variances. This appeared to be the first use of likelihood estimation for mixtures [Everitt and Hand, 1981]. However, Butler [1986] notes that there was an investigation by Newcomb [1886] of the maximum likelihood estimation of the parameters of a mixture of k univariate normal populations with known variances. His investigation could be interpreted as an application of the *EM* algorithm [Dempster, Laird, and Rubin, 1977]. Butler also found that Jeffreys [1932] had essentially used the *EM* algorithm to compute the estimates of the means in two univariate normal populations, which had known variances and which were mixed in unknown proportions.

With the advent of high speed computers, interest increased in the likelihood estimation of the parameters of mixture distributions. Hasselblad [1966, 1969] applied maximum likelihood estimation for the parameters of a mixture of k univariate normal distributions with equal variances, and then for mixtures of distributions from the exponential family. Day [1969] estimated the components of a mixture of two multivariate normal distributions with equal covariances. Wolfe [1967, 1970] used maximum likelihood estimation for the parameters of a mixture of K multivariate normal distributions with unequal covariances, and also a mixture of Bernoulli distributions. These three researchers all presented their solutions in iterative forms that could be viewed as applications of the *EM* algorithm.

For additional references on finite mixtures, see the monographs on finite mixture distributions by Everitt and Hand [1981], Titterington et al. [1985], McLachlan and Basford [1988] , the reviews by and the encyclopedia entry by Everitt [1985]

2.2 Basford's mixture-fitting programs

In their monograph on mixture models and their application to clustering McLachlan and Basford [1988] focus on the use of p-variate normal distributions for the component models and consider mainly continuous variables. Included with this book are listings of the Fortran 66 source code for four programs which estimate the parameters of normal mixture models in various situations. The program of most relevance for cluster analysis is *KMM*, which fits a mixture of multivariate normal distributions, with either arbitrary or common covariance matrices, by maximum likelihood using the *EM* algorithm. In designing the *Multimix* program, we sought to extend and modify

KMM to enhance its suitability as a general-purpose nonhierarchical clustering program. In fact *Multimix* was written from scratch, but its output was tested against that from *KMM* where possible. We are grateful to Kaye Basford for making her programs available to us in electronic form.
Some development beyond *KMM* was necessary because of three major difficulties which frustrate the application of multivariate normal mixture models to clustering. Firstly, they are not easily adapted to cope with discrete data. This is unfortunate because many real clustering problems involve both continuous and discrete variables. Secondly, they lead to models with large numbers of parameters: for example if $p = 8$ we will need to estimate 36 parameters for even a common covariance matrix, many more if they must be estimated separately for each group. A third consideration is the common occurrence of missing values in multivariate data, particularly when the observations are on humans. A variant of *Multimix* accommodates observations that are missing at random using the methodology of Little and Rubin [1987], but this is not described here.
Highly parameterized models can lead to difficulties in several ways. As discussed by McLachlan and Basford [1988, p. 11] the likelihood function of a mixture model can have singularities in a neighbourhood of which it is unbounded. Iterative methods for computing maximum likelihood estimates are drawn towards these singularities from many starting values if the model is highly parameterized. It is also common to find many local maxima in such models. Even if we find the largest of the local maxima we will often find the likelihood nearly constant in a low-dimensional set in which some of the parameters are functions of the others.

2.3 Mixtures of discrete distributions

Latent Class analysis was developed by the mathematical sociologist Paul Lazarsfeld who was interested in making more precise the relationship between underlying or latent states that are not observable, and directly observable categorical variables indicating these states.
Latent class models can be described as follows: assume the population to be made up of K groups or sub-populations $G_1, \ldots, G_K$ in proportions $\pi_1, \ldots, \pi_K$. Let $\mathbf{x}$ be the vector of responses on the p variables that we observe on each observation, where the jth variable can take on levels numbered from 1 to M_j. If the ith observation $\mathbf{x}_i$ happens to come from G_k then its probability function is given by

$$f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) = \prod_{j=1}^{p} \prod_{m} \{\lambda_{kjm} : 1 \leq m \leq M_j \quad \text{and} \quad m = x_i\}$$

where $\boldsymbol{\theta}_k$ are the parameters of the distribution of the responses in the kth subpopulation, in this case being the probabilities $\{\lambda_{kjm}\}$ that variable j takes level m, conditional on the observation belonging to group k. The overall probability function is a mixture of these conditional probability functions:

$$f(\mathbf{x}_i;\phi) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)$$

so that the latent class model is a finite mixture model. The parameter vector ϕ is made up of the π_k and the λ_{kjm} as k, j, and m take on all allowable values. Note that the π_k sum to one over k and the λ_{kjm} sum to one over m for any fixed j, k.

The original method of fitting these models, discussed at some length in Lazarsfeld and Henry [1968] for the case of binary variables, was to attempt to solve the system of equations given by equating the fitted cell probabilities to the observed cell proportions. The solution of these equations can be difficult and Latent Class analysis became much easier to use when Goodman [1974] introduced a new iterative algorithm for the maximum likelihood fitting of latent class models. It soon became clear that this algorithm was a special case of the very general *EM* algorithm.

To use latent class analysis as a clustering method the probability τ_{ik} that the ith observation comes from the kth group is first estimated by Bayes Rule from the estimated component distributions and the estimated proportions in each component. In fact these probabilities are also required in the course of the algorithm, although it is not until the algorithm has converged that we can use them for clustering. The versatility of latent class analysis as a clustering method was shown by Aitkin et al. [1981] who fitted 2-class and 3-class models to 38 binary variables describing how each of 468 teachers organised their classes, interpreting the classes as levels of a 'teaching style' factor in subsequent analyses. An even larger data set was studied by Pickering and Forbes [1984] using this method. It consisted of clinical and diagnostic information about approximately 50,000 infant births. Eleven categorical variables each having from 2 to 4 levels were used to fit models having between 1 and 6 latent classes. The analysis was feasible because only about 600 distinct response profiles actually occurred in the data. Pickering and Forbes give references to other studies using latent class methods.

Most applications of latent class analysis remain within the social sciences where the method was developed. The ability to fit latent class models is one of the capabilities of *ℓEM*, a very general program for fitting models to categorical data written by Vermunt [1997].

2.4 Everitt's model for ordinal variables

Everitt [1988] proposed incorporating binary and ordinal variables into mixture models by means of 'threshold' parameters which divide the real line into regions corresponding to outcomes of the ordinal variable. Such threshold models have been widely used for ordinal data and a brief survey is given by Zhaorong et al. [1992] where they are used in a continuous latent variable model for the comparison of 20 ternary variables representing variants of microbiological test methods. This data could also have been analyzed by latent class analysis which involves a discrete latent variable. Everitt and Mérette [1990] compare the clustering performance of Everitt's finite mixture method on four simulated data sets each having 3 continuous and 2 categorical variables, and on Fisher's iris data [Andrews and Herzberg, 1985, pp. 5-8] after two of the four variables had been categorised. They report good performance of the mixture method compared with conventional hierarchical methods. There are some severe practical limitations to the use of this method at present. Everitt proposes the use of standard optimization algorithms applied to the log-likelihood function. The computation of the log-likelihood function requires the numerical evaluation of a q-dimensional integral, where q is the number of categorical variables, and Everitt and Mérette consider no examples where $q > 2$. Their methods would be difficult to apply to the highly multivariate data sets to which cluster analysis has traditionally been applied. For these reasons *Multimix* makes no special provision for ordinal variables. Depending on the circumstances it will usually be acceptable to treat them either as categorical or continuous.

3 The *Multimix* model family

We will now describe our approach to mixture model clustering in detail. We expect the data to be in the form of an $n \times p$ matrix of observations by variables which we regard as a random sample from the distribution $f(x) = \sum \pi_k f_k(x)$, itself a finite mixture of K component distributions f_k in proportions $\pi_k \geq 0$ satisfying $\sum \pi_k = 1$. We suppose that the vector of variables $\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_p)'$ has been partitioned into $(\breve{\mathbf{x}}_1' \mid \ldots \mid \breve{\mathbf{x}}_l' \mid \ldots \mid \breve{\mathbf{x}}_L')'$. We consider component distributions that factorize $f_k(\mathbf{x}) = \prod_l f_{kl}(\breve{\mathbf{x}}_l)$, conformably with this partition. This is a weak form of *'local independence'*: within each of the K subpopulations the variables in the subvector $\breve{\mathbf{x}}_l$ are independent of the variables in $\breve{\mathbf{x}}_{l'}$ for $1 \leq l < l' \leq L$. True 'local independence' is the independence of each x_j within subpopulations. We can write the model for the ith observation as

$$f(\mathbf{x}_i; \phi) = \sum_{k=1}^{K} \pi_k \prod_{l=1}^{L} f_{kl}(\breve{\mathbf{x}}_{il}; \boldsymbol{\theta}_{kl}) \tag{2.1}$$

where $\boldsymbol{\theta}_{kl}$ consists of the parameters of the distribution f_{kl} and the π_k are the mixing proportions. This formulation includes the motivating examples of Latent Class analysis [Aitkin et al., 1981] and mixtures of multivariate normals

[McLachlan and Basford, 1988]. With one exception to be described later subvectors will usually be formed with vectors of the same type, categorical or continuous. When a subvector contains only a single variable, that variable is independent of all other variables within each subpopulation.

It is convenient to assume forms for the f_{kl}, and hence for the f_k, that belong to the exponential family. The model is then well suited for maximum likelihood estimation of its parameters by the *EM* algorithm of Dempster et al. [1977]. This approach is followed in *Multimix* with the following distributions for the $\breve{\mathbf{x}}_{kl}$:

(a) *Discrete Distribution.* Here $\breve{\mathbf{x}}_l = \{x_j\}$ is a 1-dimensional discrete random variable taking values $1, \ldots, M_j$ with probabilities $\lambda_{kl1}, \ldots, \lambda_{klM_j}$.

(b) *Multivariate Normal.* Here $\breve{\mathbf{x}}_l$ is a p_l-dimensional vector of continuous random variables with the $N_{p_l}(\boldsymbol{\mu}_{kl}, \Sigma_{kl})$ distribution.

(c) *Location Model.* Here $\breve{\mathbf{x}}_l$ is a $1+p_l$ dimensional vector of random variables with one discrete variable, x_j, and p_l continuous variables as elements. The discrete random variable takes values $1, \ldots, M_j$ with probabilities $\lambda_{kl1}, \ldots, \lambda_{klM_j}$. Conditional on the discrete variable taking value m the p_l continuous random variables have the multivariate normal distribution $N_{p_l}(\boldsymbol{\nu}_{mkl}, \Xi_{kl})$. See Krzanowski [1983] for details.

If all variables are of continuous type, then the $f(x) = \sum \pi_k f_k(x)$ will be a mixture of multivariate normal distributions. The way in which the set of variables is partitioned into subvectors determines the form of the matrix of covariance parameters in each f_k. The form is block-diagonal with a square block corresponding to each subvector. Extreme cases are the fully unstructured covariance matrix case and the diagonal covariance matrix case. Unstructured covariance matrices introduce many parameters into the model and hence should be avoided as far as possible. A reasonable strategy for fitting a mixture of multivariate normals for clustering purposes would be to begin with the local independence case (diagonal covariance matrices) and then to estimate the model parameters, assign observations to clusters and then study the within-cluster correlation matrices. Variables that are highly correlated in some of the clusters could be grouped into a subvector and the whole process repeated with the model so modified.

If all the variables are of discrete type the model is the usual Latent Class Model. In principal local independence could fail in this situation as well, although this is not often checked for. If strong within-cluster associations between two discrete variables are detected after a preliminary clustering then the two variables may be combined into a single discrete variable with a level for each cell of the two-way table (or fewer, if some cells are pooled).

The location model for a subvector in the partition is introduced in the general *Multimix* model to cope with the possibility of within-cluster associations between a discrete variable and several continuous variables. We do not expect

this facility to be needed very often in practice.
Other types of variable models are available in other mixture modelling programs to be discussed below. It is no problem in principal to add new types of attribute model from within the exponential family to extend the *Multimix* model.
As the model has been described, it is a mixture of K distributions, each of which can be seen to belong to the exponential family. It is therefore well suited for maximum likelihood estimation of its parameters by the *EM* algorithm of Dempster et al. [1977] , and the Fortran 77 program *Multimix* was written by Lynette Hunt to do this.
The 'complete data', in *EM* terminology, consists of the $n \times p$ array of observed data $\{x_{ij}\}$ and the conceptual $n \times K$ array $\{z_{ik}\}$ of class membership indicators. The indicator vectors $\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_n$ are independently and identically distributed according to a multinomial distribution generated by one draw on a population made up of K categories in proportions $\pi_1, \ldots, \pi_K$.
The complete-data specification treats the $\mathbf{z}_i$ as known leading to the log-likelihood

$$\begin{aligned} L_C(\boldsymbol{\phi}) &= \log\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\left[\pi_k^{z_{ik}}\left\{\prod_{l=1}^{L} f_{kl}(\mathbf{x}_i;\boldsymbol{\theta}_{kl})\right\}^{z_{ik}}\right]\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{z_{ik}\log\pi_k + z_{ik}\sum_{l=1}^{L}\log f_{kl}(\mathbf{x}_i;\boldsymbol{\theta}_{kl})\right\} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\pi_k + \sum_{k=1}^{K} l_k(\boldsymbol{\theta}_k) \end{aligned}$$

where

$$l_k(\boldsymbol{\theta}_k) = \sum_{i=1}^{n}\left\{z_{ik}\sum_{l=1}^{L}\log f_{kl}(\mathbf{x}_i;\boldsymbol{\theta}_{kl})\right\} = \sum_{l=1}^{L}\sum_{i=1}^{n} z_{ik}\log f_{kl}(\mathbf{x}_i;\boldsymbol{\theta}_{kl}).$$

Maximising the complete data log-likelihood $L_C(\boldsymbol{\phi})$ is equivalent to maximising $l_k(\boldsymbol{\theta}_k)$ separately for each subvector in the partition. By substituting the appropriate density for the f_{kl} Hunt [1996] deduces that the complete data sufficient statistics for the model are

1. For each class G_k the sum
$$\sum_i z_{ik};$$
2. For each class G_k, each categorical variable x_j, and each value m of x_j, the sum
$$\sum_i z_{ik}\delta_{ijm},$$

where $\delta_{ijm} = \begin{cases} 1 & \text{if } x_{ij} = m \\ 0 & \text{otherwise} \end{cases};$

3. (a) For each class G_k and each continuous variable x_j belonging to a multivariate normal subvector, the sums

$$\sum_i z_{ik} x_{ij} \qquad \text{and} \qquad \sum_i z_{ik} x_{ij}^2;$$

   (b) For each class G_k and each pair of continuous variables x_j and $x_{j'}$, $j < j'$, belonging to the same multivariate normal subvector, the sum

$$\sum_i z_{ik} x_{ij} x_{ij'};$$

4. (a) For each class G_k, each continuous variable x_j belonging to a location model subvector indexed by l and each value m of the categorical variable u_l, the sums

$$\sum_i z_{ik} w_{ilm} x_{ij} \qquad \text{and} \qquad \sum_i z_{ik} w_{ilm} x_{ij}^2;$$

   (b) For each class G_k, each pair of continuous variables x_j and $x_{j'}$, $j < j'$, belonging to the location model subvector indexed by l and each value m of the categorical variable u_l, the sum

$$\sum_i z_{ik} w_{ilm} x_{ij} x_{ij'},$$

   where $w_{ilm} = \begin{cases} 1 & \text{if } u_l = m \\ 0 & \text{otherwise} \end{cases}.$

The *EM* iteration alternates between two calculations, the E-step and the M-step. Beginning at a current value for ϕ, say $\phi^{(p)}$,the vector of all unknown parameters, the E-step requires the calculation of $Q(\phi, \phi^{(p)}) = E\{L_C(\phi) \mid X; \phi^{(p)}\}$, the expectation of the complete data log-likelihood, conditional on the observed data and the current value of the parameters. Because the complete-data sufficient statistics are linear in the unobserved z_{ik} we can calculate $Q(\phi, \phi^{(p)})$ from $L_C(\phi)$ by replacing z_{ik} with

$$\hat{\tau}_{ik} = E(z_{ik} \mid \mathbf{x}_i; \phi^{(p)}) = \frac{\pi_k^{(p)} f_k(x_i, \theta_k^{(p)})}{\sum_{k=1}^{K} \pi_k^{(p)} f_k(x_i, \theta_k^{(p)})}$$

in $L_C(\phi)$. That is, z_{ik} is replaced by the estimate of the posterior probability τ_{ik} that observation i belongs to group G_k.

At the M-Step $\boldsymbol{\phi}^{(p+1)}$ is chosen to be a value of $\boldsymbol{\phi}$ which maximises $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(p)})$ with respect to its first argument. For the *Multimix* model the elements of $\boldsymbol{\phi}^{(p+1)}$ are given by

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_{ik} \qquad \hat{\lambda}_{klm} = \frac{1}{n\hat{\pi}_k}\sum_{i,u_{il}=m}\hat{\tau}_{ik}$$

$$\hat{\boldsymbol{\mu}}_{kl} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{\tau}_{ik}\mathbf{v}_{il} \qquad \hat{\Sigma}_{kl} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{\tau}_{ik}(\mathbf{v}_{il}-\hat{\boldsymbol{\mu}}_{kl})(\mathbf{v}_{il}-\hat{\boldsymbol{\mu}}_{kl})'$$

$$\hat{\boldsymbol{\nu}}_{klm} = \frac{1}{n\hat{\pi}_k}\sum_{i,u_{il}=m}\hat{\tau}_{ik}\mathbf{v}_{ijm} \qquad \hat{\Xi}_{kl} = \frac{1}{n\hat{\pi}}\sum_{i,u_{il}=m}\hat{\tau}_{ik}(\mathbf{v}_{il}-\hat{\boldsymbol{\nu}}_{klm})(\mathbf{v}_{il}-\hat{\boldsymbol{\nu}}_{klm})'$$

for $k = 1, \ldots, K$ and $l = 1, \ldots, L$. Note that the level probabilities λ_{klm} for the categorical variables are calculated in the same way, irrespective of whether or not the discrete variable u_l belongs to a location model subvector.
The current version of the program uses a convergence criterion to cease iterating when the difference in log-likelihoods at iteration t and iteration $t-10$ is less than 0.0000001. The iteration may be started either from an initial classification or from an initial set of parameter estimates. As the number of parameters is quite large it is usually more convenient to begin with a classification.

4 Example: Byar prostate cancer data

We consider the clustering of cases on the basis of pre-trial covariates alone for the Prostate Cancer clinical trial data of Byar and Green [1980] reproduced in Andrews and Herzberg [1985, pp 261–274]. This data set was obtained from a randomized clinical trial comparing four treatments for 506 patients with prostatic cancer grouped on clinical criteria into stages 3 and 4 of the disease. As reported by Byar and Green Stage 3 represents local extension of the disease without evidence of distant metastasis, while Stage 4 represents distant metastasis as evidenced by elevated acid phosphatase, x-ray evidence, or both. We will compare the clusters obtained by *Multimix* with the clinical stages, and also consider the trial outcomes for patients in different clusters. The treatments consisted of estrogen therapy at differing rates. Daily pills containing 0.0 (placebo), 0.2, 1.0, and 5.0 mg of diethylstilbestrol were administered in the four treatments. As Byar and Green noted little difference between the effects of the first two treatments and also between the effects of the last two treatments, we will call patients in either of the first two treatments 'Untreated' and in either of the last two treatments 'Treated'.
There are twelve pre-trial covariates (Table 1) measured on each patient, seven may be taken to be continuous, four to be discrete, and one variable (SG) is

an index nearly all of whose values lie between 7 and 15, and which could be considered either discrete or continuous. We will treat SG as a continuous variable. A preliminary inspection of the data showed that the size of the primary tumour (SZ) and serum prostatic acid phosphatase (AP) were both skewed variables. These variables have therefore been transformed, SZ under a square root transformation, and AP using a logarithmic transformation, to make their distributions more symmetric. Observations that had missing values in any of the twelve pretreatment covariates were omitted from further analysis, leaving 475 out of the original 506 observations available. In fact several of the analyses to be described were also carried out using a version of the program which allows for missing observations, treating them as missing at random in the sense of Little and Rubin [1987] There was little variation from the results using only the complete observations.

We will consider the fitting of 2-class models ($K = 2$). The simplest model is the model [LInd] of complete local independence in which the component densities take the form

$$f_k(\mathbf{x}_i;\boldsymbol{\theta}_k) = \prod_{l=1}^{12} f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl}),$$

where $\boldsymbol{\theta}_{kl}$ is the parameter vector for group k, subvector l; and $k = 1,2$. Note that $f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl})$ is $N(\mu_{kl},\sigma^2_{kl})$ for each of the 8 continuous variables, and $D(\lambda_{kl1},\ldots,\lambda_{klm_l})$ for each of the 4 categorical variables.

The fitting strategy used was a form of forward selection of covariances, beginning with [LInd] and progressively adding local associations to the model by taking coarser and coarser partitions of the set of 12 variables. The modifications to the current model were determined by examining correlations, scatterplots and two-way tables within each of the two clusters formed by allocating each observation according to the current model. Table 2 summarises the results of this fitting process and a description of some of the steps follows.

When the data had been grouped into two classes following the fitting of [LInd], correlations between SBP and DBP of about 0.62 were observed within both of the classes, and these appeared to be the strongest associations. The fact that one would expect such a correlation within any naturally formed group of patients made it compelling to fit a model [BPr] in which SBP and DBP had a bivariate normal distribution within clusters. The partition of the variables for this model placed these two variables together in a subvector, the remaining subvectors being singletons. Thus [BPr] contains 2 more parameters than [LInd].

The next group of variables chosen was the triple $\{BM, WtI, HG\}$, giving a location model factor to the mixture densities as BM is dichotomous while WtI and HG are continuous. The resultant model is denoted by [3,2], referring to the size of these variable groups. Six extra parameters are introduced in this change: there are four new mean parameters, as the fitted means of WtI and HG are now specific to each level of BM within each subpopulation, and two new covariance parameters. Model [5] combines these two variable groups at a cost of introducing 12 new parameters and Model [9] has one large variable group combining BM with all 8 continuous variables. Table 2 also includes the log-likelihoods obtained. In the case of [LInd], [BPr], [3,2] and [5] these log-likelihoods were obtained from several initial configurations including random groupings of the observations; however [9] proved to be sensitive to the choice of starting configuration and the greatest log-likelihood over 4 runs is shown for this model. Convergence was usually obtained after 60 to 70 iterations although one run for model [9] reached 200 iterations without converging.

There was little difference between the group allocations determined by [LInd], [BPr], [3,2] and [5], with the allocation of only 4 patients out of 475 changing between these models. Model [9] allocations were sensitive to the initial classification and did not agree so closely with each other nor with the classifications of the more parsimonious models. Comparing the [BPr] allocation with the clinical grouping into Stages 3 and 4 of the disease we find one cluster with 252 Stage 3 and 21 Stage 4 patients and the other cluster with 21 Stage 3 and 181 Stage 4 patients.

It is of interest to examine the post-trial survival status of patients in the four Stage/ cluster combinations, which have been arrived at using pre-trial information only. This information is presented in Table 3 for the [BPr] model, and it will be noticed that while model cluster 1 and clinical Stage 3 are associated with a better chance of survival, the patterns of outcomes for the 42 patients whose model and clinical classifications conflict show that the model classifications are better indications of prognosis than the clinical criteria used. This is especially noticeable among the treated patients.

Hunt [1996] analyses this data set in more detail, also fitting 3-class and 4-class models yielding classifications with distinctively different outcome patterns suggesting that the models were detecting real features of the population. She also develops the methods of Little and Rubin [1987] for use with the model of this paper and applies this to the complete data set of 506 Prostate Cancer patients.

5 Model comparison tests

5.1 Number of components in the mixture

McLachlan and Basford [1988] devote their section 1.10 to the question of testing for the number of components in a mixture. The problem is difficult because although a model with K_1 components is nested within a model with $K_2 > K_1$ components the usual regularity conditions are not met. These conditions are required to conclude that if λ is the likelihood ratio, $-2 \log \lambda$ is asymptotically distributed as χ^2 with degrees of freedom equal to the difference in the number of parameters in the two models. In fact the asymptotic distribution may depend on the true values of the parameters of the component distributions, so there will be no general result. McLachlan and Basford discuss a number of bootstrap approaches to the problem. Feng and McCulloch [1992, 1994, 1996] have studied several aspects of this problem. In their most recent paper they recommend a bootstrap procedure. Bootstrap procedures would be very costly to apply to the clustering of data sets with many observations on many variables. Wolfe [1971] investigated the distribution of $-2\log\lambda$ when comparing nested mixtures of multivariate normal distributions and recommended treating the distribution as χ^2, but with double the nominal degrees of freedom. Banfield and Raftery [1993] developed an approximate Bayesian *approximate weight of evidence* criterion as a guide for choosing the number of components in the mixture. Wallace and Dowe [1998] use a 'minimum message length' criterion as a basis for their parameter estimation. This method unifies model selection and parameter estimation and leads to a choice for the number of components.

A specific example provides an illustration of how it may not be realistic to expect to choose a value for K on sample evidence alone. Consider the problem of estimating growth and age structure in a stock of fish from length-frequency data. Suppose that $k = 1 \ldots K$ indexes K subpopulations Π_k (age classes) of fish and that the fish in Π_k all have age t_k years, where $t_k = t_1 + k - 1$. Let π_k be the proportion of the population in Π_k and μ_k and σ_k be the mean and standard deviation of Π_k. Schnute and Fournier [1980] discuss the maximum likelihood fitting of a model of this kind where the parameters μ_k and σ_k are modelled parametrically as functions of t_k. In the fisheries application discussed by Schnute and Fournier the μ_k and σ_k tend to limiting values as k increases and the π_k vary because of annual recruitment variations but tend to diminish geometrically because of cumulative mortality, both natural and fishing. Thus as k increases the Π_k become closer together but are represented in the sample by smaller and smaller proportions. This kind of situation seems very natural but would appear to resist any form of statistical inference for the value of K. In view of the complexities of this question it seems best to regard the number of components K as a choice to be made by the modeller, in much the same way as a functional form for a distribution is selected arbitrarily. This does not mean that model comparison statistics of the $-2\log\lambda$ kind cannot be used heuristically. There remains the possibility that a small number of observations from unmodelled components will upset the fit of the model to the bulk of the data. Jorgensen [1990] discussed a number of diagnostic statistics that may be used to detect these points.

5.2 Number of within-cluster associations

In contrast to the determination of the number of classes K, the standard likelihood ratio tests for two nested models based on an approximate distribution for $-2\log\lambda$ of χ^2 with degrees of freedom equal to the difference in the number of parameters in the two models are not likely to mislead. The most troublesome regularity condition requiring checking is that the third order derivatives of $\log f$ with respect to the parameters are bounded (in a neighbourhood of the true parameter vector) by functions of the data with finite expectation [Lehmann, 1983, p. 429] . This can be shown to be the case, roughly, if the partial derivatives of the component densities f_k, with respect to all parameters and up to third order, are not too large in comparison with the mixture density f. Even for the simple case of a mixture of two bivariate normals, a model with 11 parameters we have $6^3 + 6^3$ third order derivatives of component densities to check, although many of these coincide. Checking some of these by hand suggests that all will be well unless a true variance parameter is zero or a correlation is ± 1, that is, unless a component density is degenerate. We also need neither proportion to be close to zero. A natural conjecture in the case of a finite mixture of multivariate normals is that the regularity conditions will be satisfied as long as the smallest eigenvalue of the true variance/covariance matrix for each group is not close to zero. In practical terms, the suggestion is that when a 'Reduced' model is being compared with a 'Full' model (having the same number of components, but with extra association parameters) that we may base a model comparison test on the assumption that $-2\log\lambda$ has an approximate distribution of χ^2 with degrees of freedom equal to the difference in the number of parameters in the two models when the Reduced model is operating, *unless* the fit obtained under the Reduced model has any degeneracies either in the number of components or in the form of any of the components. Similar model selection problems are considered by Dempster [1972] and Wermuth [1976a,b], but in the case of a single multivariate normal component, rather than a mixture of these. Wermuth also considers loglinear contingency table models. These authors parameterise the multivariate normal using the inverse of the covariance matrix, called the concentration matrix, rather than the covariance matrix. They test for the vanishing of a set of elements of the concentration matrix, which is equivalent to the vanishing of the corresponding set of partial correlations. As we restrict ourselves to models with block-diagonal covariance structure, and the inverses may be calculated block by block, tests involving the splitting or combining of blocks may be formulated either in terms of the covariances or the concentrations.

6 Other programs for mixture model clustering

6.1 AutoClass

AutoClass [Cheeseman and Stutz, 1996] is a Bayesian clustering program developed by Peter Cheeseman and colleagues at NASA Ames Research Center. The models fitted by AutoClass are very similar to those fitted by *Multimix*, although both programs were developed independently. Two obvious differences are

1. AutoClass has automated the process of model selection as well as that of parameter estimation but *Multimix* leaves model-specification to the user;
2. AutoClass uses Maximum Posterior estimation in place of Maximum Likelihood estimation.

In fact the first is the more crucial difference, because the *EM* algorithm at the basis of both programs accommodates both ML and MAP estimation. AutoClass compares different models by calculating an approximation to the marginal density of the observed data after the model parameters have been integrated out. In usual *EM* language the approximation used is analogous to taking observed data likelihood to be proportional to complete data likelihood with the constant of proportionality to be evaluated at the maximum likelihood estimates.
The models currently available in AutoClass for attributes within a component are as follows. Categorical attributes are modelled by general discrete distributions (multi-category Bernoulli) as in *Multimix*. Continuous attributes may be taken to have uniform or normal distributions, possibly after transformation. Poisson distributions are available for count attributes. Cheeseman and Stutz [1996] report that von Mises-Fisher distributions for circular and spherical attributes are under development. At present it appears that AutoClass does not offer facilities for modelling within cluster dependencies, that is, all models assume within-cluster independence of attributes. Missing values are treated as a special kind of value in some attribute models, but there has been no implementation of the Little and Rubin [1987] methodology for data missing at random.
Cheeseman and Stutz claim that the AutoClass method of model comparison introduces an 'Occam factor' which penalizes overfitting. However Edwards and Dowe [1998] describe the Minimum Message Length (see below) fitting of a model that combined a continuous latent factor with a number of classes to a set of 5425 infrared spectra from astronomical point sources. Edwards and Dowe found 12 classes where AutoClass had found 77 [Goebel et al., 1989]. It is not clear whether the difference in the number of classes in the fit is due to the explicit penalty on overfitting built into the Minimum Message Length

criterion or whether it is the introduction of the continuous factor which is responsible.

6.2 Snob

Snob [Wallace and Dowe, 1998] is a clustering program developed by Chris Wallace and co-workers at the Monash University Department of Computer Science, beginning in the late sixties. [Wallace and Boulton, 1968]. Snob has a home page at `http://www.cs.monash.edu.au/~dld/Snob.html`. Snob is a mixture model similar in structure to AutoClass and offering local independence models based on discrete, Normal, Poisson and von Mises distributions. In fact Snob is the older program. A novel feature of Snob is that inference is by the principal of Minimum Message Length [Wallace and Freeman, 1987]. This form of inference takes discrete variables as fundamental and seeks to minimise the negative logarithm of the probability of the model and parameter values plus the negative logarithm of the probability of the data given the model and parameter values. A continuous analogue of this estimation principle is similar to Maximum Posterior estimation (MAP) but introduces an additional factor of $(F(\boldsymbol{\theta}))^{-\frac{1}{2}}$ to the prior, where $F(\boldsymbol{\theta})$ is the determinant of the Fisher information matrix at the parameter vector $\boldsymbol{\theta}$.
In contrast to *Multimix*, where the user must specify the number of classes, Snob selects the number of classes automatically using the Minimum Message Length criterion. Thus the MML criterion is used for all aspects of model selection and parameter estimation in the Snob approach.

6.3 Mclust

Banfield and Raftery [1993] have developed the classification likelihood approach of Scott and Symons further to introduce a controlled amount of flexibility to criterion-based cluster analysis for continuous data. Although this approach suffers from the disadvantages mentioned in Section 1.2, it does lead similar optimization problems to those faced in traditional cluster analysis, and hence the model fitting may be done by algorithms similar to those used to solve those problems. Wallace and Dowe [1998] point out that in the case of a substantially overlapping pair of normal distributions having equal abundance and common σ this kind of estimation is likely to overestimate the difference in means and underestimate σ.
Bamfield and Raftery characterize the dispersion matrices of multivariate normal clusters by their *orientation*, *size*, and *shape*. They mainly consider models where the shape is the same in each component of the mixture, but orientation and size are permitted to vary. They also consider an approach to robustifying cluster analysis by allowing a very dispersed 'noise' component in addition to the multivariate normal components.

A Fortran program called *Mclust* has been written by Chris Fraley to fit these models and others. It is available from StatLib either as a Fortran program or as an S-PLUS function. Although criterion-based, rather than being based on a distance matrix, *Mclust* is written to proceed initially as an agglomerative hierarchical program. However once the number of clusters has been determined by the user *Mclust* can proceed by reallocating points to seek a minimum of the criterion in a fashion similar to the k-means algorithm of Hartigan [1975]. In recent versions of S-PLUS *Mclust* now forms the core of the clustering functions provided.

7 The place of *Multimix* in mixture modelling

The brief survey of other related programs helps to clarify the role of *Multimix* as a mixture modelling tool. In contrast to Snob and AutoClass it automates only parameter estimation, leaving model selection to the control of the user. It appears to be unique in offering a maximum likelihood approach to a class of models extending mixtures of multivariate normals and latent class models. (Although it is possible that AutoClass and Snob might be coaxed into producing similar output for at least some models by appropriate prior specification and the switching off of their model search facilities).

A natural further development for *Multimix* would be to introduce new types of attribute distribution such as the Poisson and circular von Mises distributions. To the extent that robust estimation is appropriate for a particular dataset it seems that it would be better to add a very small proportion of a highly dispersed component to the mixture than to follow Banfield and Raftery [1993] in modifying the likelihood criterion to gain robustness.

There are no present plans to automate model selection in *Multimix*, but it must be acknowledged that more needs to be done in the way of graphical diagnostic output to assist the user with the refinement of the models. Eventually some form of automation of model selection will be necessary if *Multimix* is to be used on extremely large data sets, but we would feel happier about adopting any proposal for model selection if we could compare it with human-driven procedures over a range of datasets.

The availability of the four programs AutoClass, Mclust, *Multimix* and Snob offering similar ranges of models but using different inferential principles provides an opportunity to learn more about the strengths and weaknesses of these principles in the practical data analysis context of large multivariate data sets. Currently *Multimix* is available as Fortran 77 source code from the URL `ftp://ftp.math.waikato.ac.nz/pub/maj/`. Some documentation, data sets and auxiliary programs are available at the same location.

References

M. Aitkin, D. Anderson, and J. Hinde. Statistical modelling of data on teaching styles. *J. Roy. Statist. Soc. A*, 144:419–461, 1981.

M. R. C. Anderberg. *Cluster analysis for applications.* Academic Press, New York, 1973.

D. A. Andrews and A. M. Herzberg. *Data: a collection of problems from many fields for the student and research worker.* Springer series in statistics. Springer-Verlag, New York, 1985.

J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussion clustering. *Biometrics*, 49:803–821, 1993.

D. A. Binder. Bayesian cluster analysis. *Biometrika*, 65:31–38, 1978.

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

R. W. Butler. Predictive likelihood inference with applications (with discussion). *J. R. Statist. Soc. B*, 48:1–38, 1986.

D. P. Byar and S. B. Green. The choice of treatment for cancer patients based on covariate information : application to prostate cancer. *Bull. Cancer (Paris)*, 67:477–490, 1980.

P. Cheeseman and J. Stutz. Bayesian Classification (AutoClass): Theory and Results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180, Cambridge, MA, 1996. AAAI Press/MIT Press.

R. M. Cormack. A review of classification (with discussion). *J. Roy. Statist. Soc.* B, 134:321–367, 1971.

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

A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm (with discussion). *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

R. T. Edwards and D. L. Dowe. Single factor analysis in MML mixture modelling. In X. Wu, R. Kotagiri, and K. B. Korb, editors, *Research and Development in Knowledge Discovery and Data Mining*, volume 1394 of *Lecture notes in Artificial Intelligence*, pages 96–109, Berlin, 1998. Springer.

B. S. Everitt. *Cluster analysis, Second edition.* Halsted, New York, 1980.

B. S. Everitt. Mixture distributions. In S. Kotz, N. L. Johnson, and C. B. Read, editors, *Encyclopedia of statistical sciences*, volume 5, pages 559–569, New York, 1985. Wiley.

B. S. Everitt. A finite mixture model for the clustering of mixed-mode data. *Statist. Prob. Letters*, 6:305–309, 1988.

B. S. Everitt and D.J. Hand. *Finite mixture of distributions.* Chapman and Hall, London, 1981.

B. S. Everitt and C. Mérette. The clustering of mixed-mode data: a comparison of possible approaches. *Journal of Applied Statistics*, 17:283–297, 1990.

Z. D. Feng and C. E. McCulloch. Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics and Probability Letters*, 13:325–332, 1992.

Z. D. Feng and C. E. McCulloch. On the likelihood ration test statistics for the number of components in a normal mixture with unequal variance. *Biometrics*, 50:1158–1162, 1994.

Z. D. Feng and C. E. McCulloch. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society, Series B*, 58:609–617, 1996.

H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *J. Amer. Stat. Ass.*, 62:1152–1178, 1967.

J. Goebel, K. Volk, H. Walker, F. Gerbault, P. Cheeseman, M. Self, J. Stutz, and W. Taylor. A bayesian classification of the iras lrs atlas. *Astron. Astrophys.*, 225:L5–L8, 1989.

L. A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231, 1974.

A. D. Gordon. *Classification*, volume 16 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, London, 1981.

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

J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64, 1969.

D. J. Hand. *Discrimination and classification.* Wiley, London, 1981.

J. A. Hartigan. *Clustering algorithms.* Wiley, New York, 1975.

V. Hasselblad. Estimation of parameters for a mixture of normal distributions. *Technometrics*, 8:431–444, 1966.

V. Hasselblad. Estimation of finite mixtures of distributions from the exponential family. *J. Amer. Statist. Soc.*, 64:1459–1471, 1969.

L. A. Hunt. *Clustering using Finite Mixture Models.* PhD thesis, University of Waikato, 1996.

M. James. *Classification algorithms.* Collins, London, 1985.

N. Jardine and R. Sibson. *Mathematical taxonomy.* Wiley, London, 1971.

H. Jeffreys. An alternative to the rejection of observations. *Proc. Roy. Soc. London A*, 137:78–87, 1932.

M. A. Jorgensen. Influence based diagnostics for finite mixture models. *Biometrics*, 46:1047–1058, 1990.

W. J. Krzanowski. Distance between populations using mixed continuous and categorical variables. *Biometrika*, 70:235–243, 1983.

G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies. i. hierarchical systems. *Computer Journal*, 9:373–380, 1967.

P. F. Lazarsfeld and N. W. Henry. *Latent Structure Analysis.* Houghton Mifflin, Boston, 1968.

E. H. Lehmann. *Theory of point estimation.* Wiley, New York, 1983.

R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data.* Wiley, New York, 1987.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Academic Press, London, 1979.

F. H. C. Marriot. Separating mixtures of normal distributions. *Biometrics*, 31:767–769, 1975.

G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

G. J. McLachlan and K. E. Basford. *Mixture Models : inference and applications to clustering.* Dekker, New York, 1988.

S. Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *Amer. J. Math.*, 8:343–366, 1886.

K. Pearson. Contribution to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. A*, 185:71–110, 1894.

R.M. Pickering and J. F. Forbes. A classification of Scottish infants using latent class analysis. *Statistics in Medicine*, 3:249–259, 1984.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *J. Roy. Statist. Soc. B*, 10:159–203, 1948.

J. Schnute and D. Fournier. A new approach to length-frequency analysis: growth structure. *Canadian Journal of Fisheries and Aquatic Sciences*, 37: 1337–1351, 1980.

S. L. Sclove. Population mixture models and clustering algorithms. *Commun. Statist.-Theor. Meth. A*, 6:417–434, 1977.

A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

M. J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.

W. Y. Tan and W. C. Chan. Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *J. Amer. Statist. Soc.*, 67:702–708, 1972.

D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York, 1985.

J. K. Vermunt. ℓEM. Technical report, Tilburg University, 1997.

C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.

C. S. Wallace and D. L. Dowe. MML mixture modelling of multi-state, Poisson, von Mises circular and Gaussian distributions. In L. Billard and N. I. Fisher, editors, *Proceedings of the 28th Symposium on the Interface*, volume 28 of *Computing Science and Statistics*, pages 608–613, Fairfax Station, VA, 1998. Interface Foundation of North America.

C. S. Wallace and P. R. Freeman. Estimation and inference by Compact Coding. *J. Roy. Statist. Soc. B*, 49:223–265, 1987.

N. Wermuth. Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 32:95–108, 1976a.

N. Wermuth. Model search among multiplicative models. *Biometrics*, 32: 253–263, 1976b.

J. H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioural Research*, 5:329–350, 1970.

J.H. Wolfe. NORMIX: Computations for estimating the parameters of multivariate normal mixtures of distributions. Research Memo SRM 68-2, U.S. Naval Personnel Research Activity, San Diego, 1967.

J.H. Wolfe. A monte carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Bulletin STB 72-2, U.S. Naval Personnel and Training Research Laboratory, San Diego, 1971.

J. Zhaorong, C. A. McGilchrist, and M. A. Jorgensen. Mixed model discrete regression. *Biometrical Journal*, 34:691–700, 1992.