Clustering using Finite Mixture Models

A thesis presented to the
University of Waikato
in fulfillment of the thesis requirement
for the degree of

Doctor of Philosophy

by

Lynette Ann Peach Hunt

University of Waikato
13 February 1996

Abstract

This thesis introduces a class of multivariate mixture models that includes latent class models and mixtures of multivariate normal distributions as special cases. Like latent class models, these models make free use of local independence to reduce the number of parameters in the model and to lead to descriptions of clusters that are easily understood. Provision is made for the introduction of within cluster associations between the variables.

Discrete, multivariate normal and location model distributions are the 'atoms' with which the models are built, but where more is known about the nature of the distributions in sub-populations other types of distributions could be used in place of these.

We use the EM algorithm to find the maximum likelihood estimates of the model parameters, however the emphasis is less on parameter estimation than on the use of the estimated component distributions to cluster the data. We implement the approach of multivariate mixture models with a Fortran 77 program. The program is used to fit models to several data sets, including a large medical data set. Analysis of the resulting clusters shows that sensible clusters have been achieved.

The thesis shows how our ability to analyse data using multivariate mixture models can be extended to include the facility to handle situations where data are missing at random in the sense of Rubin (1976). The program written for this thesis incorporates this facility. The scope of the methods proposed is illustrated by clustering several data sets.

Acknowledgements

Special thanks are due to my supervisor, Doctor Murray Jorgensen, for his guidance and support.

I would also like to acknowledge the support from my colleagues, and the help from Dr Wayne Schou in getting me started with TEX.

Finally, I would like to acknowledge the help given to me by my family, especially that from my two big girls, Rachael and Lorene, who became 'latch-key kids' to enable me to complete this thesis, and that from my mother, who made life easier at home for me. Thanks also to my six year old daughter Kelly, who put up with having no 'mum' at home for many nights and weekends.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Outline of Chapter 1

In this chapter, we describe the current methods for grouping data. Section 1.2 gives a brief overview of cluster analysis and points out the disadvantages of using this technique to group data. Section 1.3 describes an approach to clustering that uses a finite mixture density as its model. The iterative procedure that can be used for finding the maximum likelihood estimates in the mixture model is described in section 1.4, and section 1.5 looks at the problem of deciding the number of components to be fitted in the mixture. Section 1.6 gives an overview of latent class models and section 1.7 looks at the normal mixture model. Section 1.8 describes the approach of Everitt (1988) for clustering data sets containing both categorical and continuous variables, and section 1.9 looks at a model which has been used for data sets with categorical and continuous variables.

1.2 Clustering

The classification of objects into groups such that objects within a group are similar to each other in some way is an activity that man has performed since early times. These methods have been used in many fields such as archaeology, psychology, medicine and market research. The development of the computer has led to an increase in the use of automated numerical methods of classification. Much of the work has been done without knowledge of related work in other disciplines (Gordon (1981)).

Cluster analysis attempts to identify any possible tendency of the data to clump together to form groups. There is no prior information regarding the underlying group structure, and the goal is to partition the data into groups such that members of a group are as similar as possible.

This contrasts with another classification procedure, discriminant analysis, where the existence of a set of relevant groups are known *a priori* and the aim is to produce a rule for classifying new individuals. Two types of multivariate observations are present in this situation. The first, the training set, are those observations whose membership in a specific one of K given groups is known, and the second, referred to as test samples, are observations for which group membership is unknown, and which have to be assigned to one of the K groups. A comprehensive

review of discriminant analysis has been given by Hand (1981), and McLachlan (1992).

There are many different methods of cluster analysis. These methods can be broadly categorised as hierarchical or non-hierarchical. Clustering using hierarchical methods, is generally obtained through two types of algorithm: (i) agglomerative algorithms in which there is a successive pooling of subsets of the set of objects, and (ii) divisive algorithms in which there is a successive partitioning of the set of objects. 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.

Divisive methods proceed with all n observations considered as a single cluster. This single cluster is split into two sub clusters. One of these sub clusters is then split into two further sub clusters and so on. The decision about which cluster to split and how to split it can be based on variables considered one at a time (monothetic techniques), or on all variables considered simultaneously (polythetic techniques). See Hand (1981) for further details.

The agglomerative methods of hierarchical cluster analysis techniques include some of the oldest and most popular methods of cluster analysis. The agglomerative methods proceed with n clusters, each consisting of one observation. The next step is to merge two clusters to yield $n-1$ clusters. The two clusters to be merged are chosen by studying the similarity matrix of inter-cluster similarities. The number of clusters is then reduced by one by fusing the two clusters considered to be the most similar to each other. This is repeated until the required number of groups is present or until the final stage in which all n observations are in one cluster.

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 dendrogram, also known as a tree diagram.

1.2.1 Some hierarchical clustering techniques

Most of the hierarchical clustering techniques can be implemented with the data represented by a matrix of proximities $\mathbf{D} = (d_{ij})$, where d_{ij} is the proximity of observations i and j. The proximity d_{ij}, can either be a similarity or a dissimilarity

measure. 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 $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$, whilst the Mahalanobis distance is defined as $d_{ij}^2 = [(\mathbf{x}_i - \mathbf{x}_j)'\hat{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)]^{1/2}$, where $\hat{\Sigma}^{-1}$ is the within cluster covariance matrix. Further details on the properties of these distances and other distance measures have been given by for example, Mardia, Kent and Bibby (1979), and Gordon (1981).

Consider the following two approaches to cluster analysis. With single linkage (nearest neighbour) clustering, the distance between two clusters is defined as the distance between their two nearest neighbours. That is,

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

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 being formed as once links are formed, they cannot be broken. As the clusters have no nuclei, a 'chaining' effect results. Jardine and Sibson (1971) show that single linkage clustering is the only technique to satisfy various analytical properties.

Whereas with complete linkage (farthest neighbour) clustering, the distance between two clusters is defined as the distance between their two furthest neighbours. That is,

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

where d_{AB} and d_{ij} are dissimilarities between clusters A and B, and observations i and j respectively. This method tends to produce compact clusters with no chaining effect, but it does not necessarily find all groups where the within-group distances are less than some value.

These techniques use different distance measures to cluster data, and for a given data set may produce quite different clusters.

Lance and Williams (1967) have given a general agglomerative algorithm with which many of the common hierarchical 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. A table of the coefficients for different techniques has been given by Gordon (1981), and Cormack (1971).

1.2.2 Non hierarchical techniques

Non hierarchical 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 non hierarchical 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. With this strategy the structure of the groups is optimised and the groups are thus made as homogeneous as possible. 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 example, Everitt (1980) and Hand (1981).

The k means algorithm described in 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 of the observations in that cluster to that point.

The classification likelihood approach is a non hierarchical technique that uses a likelihood based approach to clustering. 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

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

Let $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and $\boldsymbol{\phi} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. Maximisation of $L_C(\mathbf{z}, \boldsymbol{\phi})$, the loglikelihood 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 ϕ. 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 John (1970), Sclove (1977), Scott and Symons (1971) and Symons (1981). Unfortunately with this procedure, the z_{ij} increase in number with the number of observations, and the maximum likelihood estimates are not necessarily consistent.

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. Marriott (1975) pointed out that the maximum 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. Under this approach, the $\mathbf{z}_i$ are treated as if they are parameters, rather than as missing random variables. Symons (1981) and Binder (1978) give Bayesian versions of this method, but we will not discuss this approach in this thesis.

There is a vast quantity of literature available on cluster analysis, and it is not possible to list all references. For comprehensive reviews of clustering techniques see for example, Cormack (1971), Everitt (1980), Jardine and Sibson (1971), and Gordon (1981). For clustering algorithms see Hartigan (1975) and James (1985).

1.2.3 The number of clusters

In the discussions so far, we have assumed that there are K clusters in the data set, but no mention has been made of how we evaluate this figure K. Prior knowledge concerning the number of clusters reduces the complexity of a cluster analysis. However, the situation often occurs where we have a sample of data in which the number of clusters K is unknown. Finding the number of clusters present in the data then becomes part of the clustering problem. A general approach is to compare some criterion evaluated for different levels of K. Marriott (1971) suggested that the value of K for which $K^2 \mid \Sigma \mid$ is a minimum will give the optimum number of clusters, where Σ is the within cluster covariance matrix.

With hierarchical clustering, the number of clusters is obtained by selecting one of the clusterings in the nested sequence of groupings displayed in the dendrogram. The most common method used is to examine the dendrogram for large changes in the distance or dissimilarity between adjacent fusion levels. A large change

when going from K_1 to $K_1 - 1$ groups might be indicative of K_1 groups. However, this procedure is subjective, and two users of this approach may recommend a different number of groups from examination of the same dendrogram. Mojena (1977) suggests a procedure based on the relative sizes of the different fusion levels. Milligan and Cooper (1985) discuss various rules for selecting the number of groups from the dendrogram. These tests can assist in assessing the number of clusters present in the data.

1.2.4 The disadvantages of cluster analysis

When a cluster analysis is performed there are some inescapable drawbacks. Any randomness in the data is not reflected in the grouping and with some common algorithms, a small perturbation in the data can lead to quite different clusters being formed. All clustering methods make implicit assumptions about the type of structure present in the data. Unless there is a specific reason for using a particular method, it must be ensured that the conclusions are not just an artefact of the method used. For example, hierarchical techniques impose a hierarchical structure on the data where it may not be present. Group average, complete linkage and Ward's method often find spherical clusters even when it appears that the data contain clusters of other shapes (Everitt and Dunn, 1991). Single linkage cluster analysis tends to find rod shaped clusters especially when there are intermediates present between the clusters.

Optimisation methods tend to find spherical and elliptical clusters. If the data contains other shapes, they may not be found by using these methods.

Considerable care must be taken when using cluster analysis to avoid misleading solutions, as clustering techniques will generate a set of clusters even when applied to random unclustered data. If there is strong structure present in the data, one hopes that it is picked up by most techniques. Hand (1981) recommends applying different clustering techniques to the data set as one way of verifying the structure.

Optimisation techniques usually require large amounts of computer time as they try various allocations of points to clusters. The underlying assumptions about the distribution such as having a common covariance structure for the clusters, are important for these techniques.

Hierarchical techniques operate on the matrix of proximities between the observations rather than the observations themselves, and thus the use of these methods entails a loss of information (Mardia, Kent and Bibby, 1979). Using a different clustering algorithm can result in a different grouping structure. Further, experience with real mixed populations shows that they are often substantially overlapping, whereas by design cluster analysis will tend to come up with compact non

overlapping clusters.

Another disadvantage with cluster analysis is that after deciding on the clustering algorithm to be used, the user still has to make more decisions such as deciding on the distance measure to use, whether or not to standardise the variables - different sets of weights on the variables can lead to completely different clusters. In the next section, we will consider an approach to clustering that uses a finite mixture density as its model. Advantages of this model based approach include completely dispensing with the need to decide on which similarity measure is appropriate and which clustering algorithm to use.

1.3 The Finite Mixture model

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 have 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 clustering.

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 is an example of a non hierarchical clustering technique. As pointed out by Aitkin, Anderson and Hinde (1981), "clustering methods based on mixture models allow estimation and hypothesis testing within the framework of standard statistical theory".

1.3.1 History

The finite mixture problem has quite a lengthy history. Karl Pearson in 1884 put forward a solution in the case of a mixture of two univariate distributions with unequal variances using the method of moments. This was a difficult problem and involved the solution of a ninth degree polynomial equation. Later investigation, for example, Tan and Chan (1972), showed that likelihood estimation was superior to the method of moments for this problem.

Maximum likelihood estimation for the parameters in mixture distributions was suggested in 1948 by Rao, who used Fisher's method of scoring for the estima-

tion 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) pointed out that there was an investigation by Newcomb in 1886, for the maximum likelihood estimation of the parameters of a mixture of k univariate normal populations with known variances. This investigation could be interpreted as an application of the EM algorithm of Dempster, Laird and Rubin (1977). Butler also found that Jeffreys in 1932 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 an iterative form that corresponded to applications of the EM algorithm of Dempster, Laird and Rubin (1977). Further details on the EM algorithm can be found in section 1.4.

There is an extensive literature on the finite mixture problem, and it is not possible to list all references. For additional references on finite mixtures, see the monographs on finite mixture distributions by Everitt and Hand (1981), Titterington, Smith and Makov (1985), and McLachlan and Basford (1988), the reviews by Redner and Walker (1984) and encyclopedia entries by Everitt (1985).

1.3.2 The mixture model

Suppose that p attributes are measured on n observations. We regard the observations to be clustered as a random sample of a mixture from a finite number, K, of populations in some unknown proportions, $\pi_1, \ldots, \pi_K$, respectively, where $\pi_k \geq 0$ and $\sum \pi_k = 1$. The probability density function of an observation $\mathbf{x}$ can be represented in the finite mixture form

$$f(\mathbf{x}_i; \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \tag{1.1}$$

where $f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)$ is the density of $\mathbf{x}_i$ in group k, $\boldsymbol{\theta}_k$ denotes the vector of unknown parameters for group k, and $\boldsymbol{\phi} = (\boldsymbol{\pi}', \boldsymbol{\theta}')'$, the vector of all unknown parameters.

Note that we have overparameterized here; we only need $K-1$ of the π_k since the sum of the π_k over the K groups equals one.

The likelihood function for $\boldsymbol{\phi}$ under (1.1), is given by

$$l = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right\}.$$

Let $\hat{\boldsymbol{\phi}}$ be the maximum likelihood estimate of $\boldsymbol{\phi}$. Then each observation, $\mathbf{x}_i$, can be allocated to group k on the basis of the estimated posterior probabilities. The k^{th} mixing proportion π_k, can be viewed as the prior probability that the observation belongs to group k. Under the mixture model, the posterior probability that observation $\mathbf{x}_i$, belongs to group k, is given by

$$\begin{aligned} \tau_k(\mathbf{x}_i; \boldsymbol{\phi}) &= pr(\text{observation } i \in \text{ group } k \mid \mathbf{x}_i; \boldsymbol{\phi}) \\ &= \frac{\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)} \end{aligned}$$

for $k = 1, \ldots, K$. Note that the use of 'posterior' and 'prior' is not related to the method used to estimate the parameters $\boldsymbol{\phi}$.

If $\boldsymbol{\phi} = (\boldsymbol{\pi}', \boldsymbol{\theta}')'$ is estimated by $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\pi}}', \hat{\boldsymbol{\theta}}')'$, then the observations can be partitioned into K non overlapping clusters by assigning each observation to the group to which it has the highest posterior probability of belonging. That is, $\mathbf{x}_i$ is assigned to group k, if

$$\tau_k(\mathbf{x}_i; \hat{\boldsymbol{\phi}}) > \tau_{k'}(\mathbf{x}_i; \hat{\boldsymbol{\phi}}) \text{ for } k = 1, \ldots, K;\ k \neq k'.$$

If the parameters $\boldsymbol{\phi}$ are known, this allocation rule would be the optimal or Bayes rule which minimises the overall error rate (Anderson, 1958). For convenience, $\tau_k(\mathbf{x}_i; \hat{\boldsymbol{\phi}})$ will be denoted as $\hat{z}_{ik}$.

The mixture model, (1.1), can be fitted parametrically, using the method of maximum likelihood. In order to estimate the parameters in $\boldsymbol{\phi}$, from the data, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, it is necessary that the parameters be identifiable. The mixture distribution in (1.1) is identifiable if there exists a one-to-one correspondence between the mixing distribution and the resulting mixture (Everitt, 1985). For further details on identifiability, see McLachlan and Basford (1988), and Titterington, Smith and Makov (1985).

In some instances, the loglikelihood $L(\phi)$ is unbounded under the mixture model, and so the maximum likelihood estimator of $\boldsymbol{\phi}$ does not exist. These singularities

occur at certain points on the boundary of the parameter space, and a sensible local maxima can be found for most examples, provided we keep away from these boundaries (Titterington, Smith and Makov, 1985). The loglikelihood function may also have local maxima. The problem then arises as to whether to accept a given local maximum, or to go on and search for others. If the root of the likelihood equation is not unique, an obvious choice is the root corresponding to the largest of the local maxima, assuming that these have all been found. Redner and Walker (1984) have given, for identifiable mixtures, the regularity conditions that must be satisfied for a sequence of roots of the likelihood equation $\partial L/\partial \phi = 0$ to exist and have the properties of consistency, efficiency and asymptotic normality. McLachlan and Basford (1988) maintain that the form of the regularity conditions suggest that they hold for many parametric families.

In the next section, we describe the EM algorithm of Dempster, Laird and Rubin (1977), an iterative procedure that will be used to find the maximum likelihood estimates in the finite mixture model.

1.4 The EM Algorithm

The EM algorithm is a general iterative procedure for computing the maximum likelihood estimates when the data can be viewed as incomplete. The EM algorithm is simple conceptually. In simple cases, the algorithm can be informally described as follows: First, formulate the problem such that if the missing data were observed, the maximum likelihood estimate of the parameter θ, would be easy to find. Then (1) fill in the missing values, (2) estimate the parameters, and (3) use the new parameter estimates to re-estimate the missing values, (4) re-estimate the parameters and so forth, iterating until convergence. In most cases, the EM algorithm is more complicated. Generally, missing sufficient statistics need to be estimated, and the loglikelihood needs to be estimated at each iteration of the algorithm.

The term, 'EM algorithm', was first applied by Dempster, Laird and Rubin (1977). In this paper, they identified the full generality of the algorithm and proved general results about its behaviour. The literature was reviewed and a wide range of examples was provided. The algorithm was termed 'EM', signifying Expectation - Maximisation, as each iteration consists of an expectation step (E step), and a maximisation step (M step).

Previous researchers, for example, Dempster, Laird and Rubin (1977), and Little and Rubin (1987), found that the EM algorithm had a long history of application in special cases prior to 1977, with the earliest EM application found being a medical application by Kendrick in 1926. Orchard and Woodbury (1972) give a general

approach to likelihood estimation for incomplete data, and consider a mixture of multivariate normal distributions, where the group assignment is regarded as missing information. Sundberg (1974) considered properties of the general likelihood equation, and Beale and Little (1975) further developed the theory for analysing multivariate normal data with missing values. Rubin (1991) presents a review of the ideas and basic theme of EM, and discusses some algorithmic techniques such as multiple imputation, that combine simulation techniques with complete data methods to address problems that are difficult or impossible to solve using the EM algorithm.

The EM algorithm has several important properties. An advantage of the algorithm is that it can be shown to converge reliably, as each iteration of the algorithm, increases the loglikelihood, $L(\boldsymbol{\phi} \mid \mathbf{x})$, where $\boldsymbol{\phi}$ is the parameter to be estimated. Thus,

$$L\left(\boldsymbol{\phi}^{(t+1)} \mid \mathbf{x}\right) \geq L\left(\boldsymbol{\phi}^{(t)} \mid \mathbf{x}\right),$$

where $L(\boldsymbol{\phi}^{(t)} \mid \mathbf{x})$ is the loglikelihood at iteration t. See Dempster, Laird and Rubin (1977), and Rubin (1991) for further details on the properties of the EM algorithm. Wu (1983) gives a detailed account of the convergence properties of the EM algorithm and corrects some errors made by Dempster, Laird and Rubin.

The main disadvantage of the EM algorithm is that the rate of convergence can be slow. The greater the proportion of missing data, the slower the rate of convergence. Specifically, Dempster, Laird and Rubin (1977) show that when the eigenvalues of the jacobian matrix of the update are all less than one, the largest of these eigenvalues gives the rate of convergence of the algorithm. That is, the rate of convergence depends on the relative sizes of the information in the observed data about $\boldsymbol{\phi}$, and the expected or Fishers information in the unobserved data about $\boldsymbol{\phi}$. Another disadvantage is that the EM algorithm computes the maximum likelihood estimate of the parameters but does not directly give the observed information matrix that may be required to give the standard errors of the parameters. Louis (1982) presents a procedure for extracting the observed information matrix when using the EM algorithm, and a method for speeding up the convergence of the algorithm. Further details on the observed information matrix for mixture models are given by McLachlan and Basford (1988, section 1.9). Meilijson (1989) presents a unification of EM methodology and Newton methods for speeding up the algorithm. Methods for speeding up the convergence of the EM algorithm are an area of current research.

There are other general iterative procedures such as the Newton Raphson algorithm and the method of scoring, that could also be used for the solution of the likelihood equations. It can be shown that if the loglikelihood function under the

mixture model is concave and unimodal, the sequence of estimates of ϕ produced by the Newton Raphson algorithm converges quadratically near the maximum. However this algorithm does not always increase the likelihood as it can move to a local minimum, and therefore choice of starting values is more important with Newton's method than with the EM algorithm. At each iteration, the Newton Raphson algorithm requires the calculation and/or storage of the second derivative matrix of the loglikelihood. It can be seen that use of this algorithm can be infeasible in many problems, especially clustering problems with large numbers of observations and variables. The method of scoring involves the calculation and inversion of Fishers information matrix. The computations per iteration can be fairly large. This method also does not always increase the likelihood. See Redner and Walker (1984), and Titterington, Smith and Makov (1985, pages 88-89) for further details on both these algorithms.

1.4.1 EM for exponential families

Suppose that the complete data $\mathbf{x}$ have a distribution from an exponential family defined by

$$f(\mathbf{x} \mid \phi) = b(\mathbf{x}) \exp(s(\mathbf{x})\phi)/a(\phi),$$

where ϕ is a $(p \times 1)$ parameter vector, $s(\mathbf{x})$ is a $(1 \times p)$ vector of complete data sufficient statistics, and $b(\mathbf{x})$ and $a(\phi)$ are functions of $\mathbf{x}$ and ϕ respectively.

When the EM algorithm is applied to incomplete data from the exponential family, the E step consists in estimating the complete data sufficient statistics $s(\mathbf{x})$ given by

$$s^{(t)} = E(s(\mathbf{x}) \mid \mathbf{x}_{obs}, \phi^{(t)}),$$

where $\phi^{(t)}$ denotes the current value of ϕ at iteration t, and $\mathbf{x}_{obs}$ denotes the observed part of $\mathbf{x}$.

The M step determines the new estimate $\phi^{(t+1)}$ of ϕ as the solution of the likelihood equations

$$E(s(\mathbf{x}) \mid \phi) = s^{(t)},$$

which are the likelihood equations for the complete data $\mathbf{x}$ with $s(\mathbf{x})$ replaced by $s^{(t)}$.

We see that when the complete data comes from the exponential family, the E step reduces to finding the conditional expectation of the complete data sufficient statistics. The M step is simple and involves complete data maximum likelihood estimation.

The EM algorithm for mixture density problems can be regarded as a specialisation of the general EM algorithm given by Dempster, Laird and Rubin (1977). The E step of the algorithm requires the computation of $Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}^{(t)})$, where
$Q(\boldsymbol{\phi}' \mid \boldsymbol{\phi}) = E(\log f(\mathbf{x} \mid \boldsymbol{\phi}') \mid \mathbf{x}_{(obs)}, \boldsymbol{\phi})$, this is assumed to exist for all pairs $(\boldsymbol{\phi}', \boldsymbol{\phi})$;
$f(\mathbf{x} \mid \boldsymbol{\phi})$ is the density function defined for the complete data sample space;
$\mathbf{x}_{(obs)}$ are the observed data; and $\boldsymbol{\phi}$ is the parameter to be estimated.

The M step consists of choosing $\boldsymbol{\phi}^{(t+1)}$ to be the value of $\boldsymbol{\phi}$ that maximises $Q(\boldsymbol{\phi}' \mid \boldsymbol{\phi}^{(t)})$.

Dempster, Laird and Rubin (1977) considered the estimation of the parameters of the mixture density as an estimation problem involving incomplete data, by regarding each observation as 'missing' a label indicating its component population of origin. The algorithm is applied to the mixture model using the general formulation described above. Rather than adjoining a single variable of class assignments, it is more convenient to add K indicator variables corresponding to each of the K classes.

Let the vector of indicator variables, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$, be defined by

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

where the indicator vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independently and identically distributed according to a multinomial distribution generated by one draw on a population made up of K categories in proportions $\pi_1, \ldots, \pi_K$. Further suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are conditionally independent.

Maximum likelihood estimation for the 'complete data' formed by the adjoining of 'missing data' onto the observed data, is then simple. The complete data loglikelihood for the model (1.1) can be written in the form

$$L_C(\boldsymbol{\phi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left\{ \log \pi_k + \log f_i(\mathbf{x}_i; \boldsymbol{\theta}_k) \right\}.$$

Since the complete data likelihood is linear in the z_{ik}, the E step of the algorithm requires the estimation of the z_{ik} given the data and the current vaue of the parameters. The M step separates into two maximisation problems, one involving the mixing proportions $\pi_1, \ldots, \pi_K$ and the other involving the parameters of the K distributions. For further details on EM for mixture densities, see Redner and Walker (1984, section 4) and Titterington, Smith and Makov (1985). For details on convergence properties of EM, and properties on the estimators of mixtures of densities from the exponential family, see Redner and Walker (1984, section 5).

1.4.2 The choice of starting values for the EM algorithm

The EM algorithm can be started in two ways. We can either input some initial value for ϕ, say ϕ^0, or we can execute the M step of the EM algorithm with an initial estimate of the group assignment. That is, input estimates of the posterior probabilities z_{ik}, where z_{ik} is set at 1 or 0 according to whether observation i is in group k or not.

To help select starting values for the EM algorithm, McLachlan (1988) recommends that a principal component analysis be firstly carried out on the data. Then, the two dimensional scatter plots of the principal components can be examined for the presence of clusters. Evidence of any clustering can then be reflected in the initial estimates of the posterior probabilities of group membership. With medical applications such as the cancer data investigated in Chapter 3, there is often some *a priori* information concerning a possible group structure for the patients under study. This can extend to a provisional grouping of the n unclassified patients.

As the likelihood equation for mixture models usually has multiple roots, the selection of suitable starting values for the EM algorithm is important. The EM Algorithm should also be applied from different starting values in any search of local maxima. If there are several local maxima, the problem then arises as to which root of the likelihood equation should be used. An obvious choice is the one corresponding to the largest of the local maxima, although it does not necessarily follow that the consequent estimator is consistent and asymptotically efficient (Lehmann (1980)).

1.5 Choosing the number of groups.

In many situations in practice, there is no *a priori* knowledge of the number of classes to be fitted to the model. An obvious way of approaching this problem is to use the likelihood ratio test statistic λ to test for the smallest value of K compatible with the data. However when testing for the number of components in a mixture, the usual regularity conditions do not hold for $-2\log\lambda$ to have its standard asymptotic null distribution of χ^2 with the degrees of freedom equal to the difference between the number of parameters under the full and reduced models.

For example, consider a two component mixture model where

$$f(\mathbf{x};\phi) = \pi_1 f(\mathbf{x};\theta_1) + (1-\pi_1) f(\mathbf{x};\theta_2)$$

with $0 \leq \pi_1 \leq 1$. To test for the existence of two components in the mixture, the null hypothesis $H_0 : K = 1$ is approached by testing that $\pi_1 = 1$, which falls on the boundary of the parameter space. An alternative formulation would be to test

$H_0 : \theta_1 = \theta_2$, a test that the two components are identical. This test does not correspond to a point on the boundary of the parameter space, but if $\theta_1 = \theta_2$, the mixture

$$f(\mathbf{x}; \boldsymbol{\phi}) = \pi_1 f(\mathbf{x}; \theta_1) + (1 - \pi_1) f(\mathbf{x}; \theta_1) = f(\mathbf{x}; \theta_1)$$

regardless of the value of π and the likelihood function is flat in π.

Titterington (1981), and Titterington, Smith and Makov (1985) investigated the likelihood ratio test of $H_0 : K = 1$ versus $H_a : K = 2$ for a mixture of two known densities in unknown proportions. They showed that, $-2 \log \lambda$ is asymptotically zero with probability 0.5, under the null hypothesis. The distributional result $-2 \log \lambda \sim \chi^2_{(1)}$, expected under the usual asymptotic theory did not hold under H_0. Hartigan (1985a,b) obtained the same result for the asymptotic null distribution of $-2 \log \lambda$ for a mixture of two univariate normal distributions with common covariances, known means and in unknown proportions. Further accounts of the breakdown of the regularity conditions are given for example, by Ghosh and Sen (1985), and Hartigan (1977).

The use of the likelihood ratio test for more complicated problems is more difficult. Some of the early investigations of this test used restricted numerical studies. In some investigations such as Hasselblad (1969), the likelihood ratio test was applied and the problem of the breakdown in the regularity conditions was not recognised. Aitkin and Tunnicliffe Wilson (1980) use the likelihood ratio test whilst pointing out that the asymptotic distribution may not hold satisfactorily especially with small samples, and that the χ^2 test with degrees of freedom on the number of parameters is not valid as mixtures are not regular.

To test the hypotheses, $H_0 : K = K_1$ versus $H_a : K = K_2$ for a mixture of p variate normals where $K_1 < K_2$, Wolfe (1971) suggested on the basis of a small scale simulation performed for $K = 1$ and $K = 2$, that the null distribution of $-2 \log \lambda$ could be approximated as $-2c \log \lambda \sim \chi^2_d$ where the degrees of freedom, d, is twice the difference in the number of parameters in the two hypotheses, not including the mixing proportions. The suggested value of c is $(n-1-p-0.5K_2)/n)$.

The simulations of Everitt (1981) for testing that $K = 1$ versus $K = 2$ for mixtures of normal populations with equal covariance matrices, suggest that the ratio n/p needs to be at least 5 for Wolfe's approximation to be applicable for the determination of P values. These simulations had a sample size of between 25 and 500, for p up to 10. Everitt (1981) also shows that the power of the test is small unless the component densities are well separated.

McLachlan and Basford (1988, page 24) recommend that Wolfe's modified ratio test be used as a guide to the possible number of underlying groups. They also

suggest that use also be made of the estimates of the posterior probabilities group membership in the choice of the number of groups.

Aitkin, Anderson and Hinde (1981) expressed reservations about the adequacy of the chi squared approximation for the distribution of $-2\log\lambda$, and outlined a solution to the problem by using Monte Carlo methods for latent class models. This was described by McLachan and Basford (1988) as essentially a bootstrap approach. The likelihood ratio test statistic for the test of $H_o : K = K_1$ versus $H_a : K = K_2$ can be 'bootstrapped' as follows. Proceeding under the null hypothesis, the likelihood estimates $\hat{\phi}$, are found for the data by fitting a mixture model with K_1 groups. A sample is then generated from the mixture

$$f(\mathbf{x};\hat{\phi}) = \sum_{k=1}^{K_1} \hat{\pi}_k f_k(\mathbf{x};\hat{\boldsymbol{\theta}}_k) \ ,$$

where the parameter estimates for the distribution are the likelihood estimates obtained from fitting the model to the original data. Mixture models for K_1 groups, and then K_2 groups are fitted for the sample, and $-2\log\lambda$ is computed. Another sample is generated, and the process is repeated independently T times. The replicated values of $-2\log\lambda$ calculated from the successive samples can be compared with that from the original data. Further references on this bootstrap approach are given by McLachlan and Basford (1988).

McLachlan (1987) assesses the bootstrapping of $-2\log\lambda$ for the test of a single normal density versus a mixture of two normal densities in the univariate case where both distributions have a common variance. For a sample size of 100, McLachlan found that as the number of replications of the bootstrap increased, the simulated power also increased and that the power of the test was poor unless the components of the mixture are widely separated.

For the test of a single normal density versus a mixture of two normal densities with unequal variances in the univariate case, Feng and McCulloch (1994) show that the criterion used to avoid singular solutions, affects the distribution of $-2\log\lambda$.

To avoid the failure of the standard likelihood ratio test, Aitkin and Rubin (1985) suggest an approach which places a prior distribution on the vector of mixing proportions $\boldsymbol{\pi}$. The likelihood $L(\boldsymbol{\theta} \mid \mathbf{x})$ is firstly found by integrating the likelihood $L(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{x})$ over the prior distribution of $\boldsymbol{\pi}$. The distributional vector $\boldsymbol{\theta}$ is then estimated by maximum likelihood from $L(\boldsymbol{\theta} \mid \mathbf{x})$. Tests on the number of groups are made relative to the one group model. Aitkin and Rubin (1985) maintain that the usual asymptotic distribution can be expected to apply with this proposal if the sample size is not small. However, Quinn, McLachlan and Hjorth (1987) demonstrate that even with the use of a prior distribution on the $\boldsymbol{\pi}$, the regularity

conditions usually assumed to obtain the standard asymptotic results still do not hold.

The use of Akaike's information criterion (Akaike (1974)) to assess the number of components in a mixture was proposed by Sclove (1983), and Bozdogan and Sclove (1984). Titterington(1984), and Titterington, Smith and Makov (1985) point out that this criterion relies essentially on the regularity conditions needed for $-2\log\lambda$ to have its usual asymptotic distribution under the null hypothesis.

Assessment of the number of modes is another way of approaching the problem for the number of components in a mixture. However, as bimodality can occur in a single normal distribution and conversely, mixtures can be unimodal, care must be taken. (See for example, Everitt and Hand (1981) and Titterington, Smith and Makov (1985)).

The main problem is the lack of identifiability of the parameters even when the class of mixtures is identifiable. See for example, Hartigan (1985a), Titterington, Smith and Makov (1985).

As can been seen, the task of identifying the number of components in a mixture by using statistical tests is still an unresolved problem. It should be noted that the question of the number of groups does not have any real meaning unless (i) the mixing proportions are greater than some value, and (ii) any two component distributions are sufficiently different in some measure. Knowledge of the subject area must be used to determine whether the clusters actually have a real meaning, or whether we have just imposed a structure on the data. The underlying structure of the groups cannot be checked as it is unknown, we estimate the parameters of the distribution and the mixing proportions.

In this thesis, we will use the likelihood ratio test merely as a guide to the possible number of underlying groups in the mixture. Another guide can be found in the estimates of the posterior probabilities of group membership. Clearly a solution where observations are clearly assigned to a particular component will be of more practical use than one in which many observations have appreciable probability of membership in each of several classes. It must be remembered, however, that real populations do overlap, and such solutions are not necessarily meaningless.

1.6 Latent Class analysis as a Clustering method

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 were not observable, and directly observable indicators of those states, the manifest variables. In latent class analysis, the observed manifest

variables are categorical and it is assumed that the underlying latent variable is also categorical. The basic assumption of the latent class model is that within any category of the latent variable, the manifest variables are independent of each other. This assumption is known as *the axiom of local independence*. The observed relationships between the manifest variables thus result from the underlying classification of the data produced by the unobserved latent variable.

Latent class models can be expressed as a finite mixture distribution as follows: we assume the population to be made up of K groups or sub-populations $G_1, \ldots, G_K$ in proportions $\pi_1, \ldots, \pi_K$, where $\pi_k \geq 0$ and $\sum \pi_k = 1$. Let $\mathbf{x}$ be the vector of responses on the p variables that we observe on each observation, where the jth variable can have levels numbered from 1 to M_j. Let λ_{kjm} be the probability that variable j takes level m in group k. Then, if the ith observation $\mathbf{x}_i$ happens to come from G_k, its probability function is given by

$$f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) = \prod_{j=1}^{p} \lambda_{kjx_{ij}}$$

where $\boldsymbol{\theta}_k$, the vector of unknown parameters of the distribution of the responses in the kth sub-population has the $\{\lambda_{kjm}\}$ as components.

The overall probability function is a mixture of these conditional probability functions:

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

The parameter vector $\boldsymbol{\phi}$ is made up of the π_k and the λ_{kjm} as k, j, and m take on all allowable values. We have overparameterized here as the π_k summed over k and the λ_{kjm} summed over m for any fixed j, k will total 1; thus we should regard, say, π_K and the λ_{kjM_j} as excluded from $\boldsymbol{\phi}$, which has $K - 1 + K\left(\sum_{j=1}^{p} M_j - p\right)$ elements.

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 known as the 'accounting equations', which equate the observed cell proportions to the predicted cell probabilities. These equations were used as the basis of parameter estimation for approximately 20 years. The solution of these equations can be difficult. Further details can be found for example, in Everitt (1984a).

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 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, the estimates of these probabilities are also required in the course of the algorithm, although it is not until the algorithm has converged that they can be used for clustering.

Everitt (1984a,b) gave a numerical example of the latent class model using the machine-design data previously analysed in 1956 by McHugh and displayed in Table 1.1. The data set arose from four machine-design subtests that were given to 137 engineers: the data were dichotomised such that '1' signifies above the subtest mean and '0' is below the subtest mean.

Table 1.1

Machine design data from Everitt (1984a)

Response Pattern				Frequency		
Variable				Observed	Expected	
1	2	3	4		2-class	3-class
1	1	1	1	23	20.29	19.73
0	1	1	1	8	10.35	10.83
1	0	1	1	6	9.27	9.33
1	1	0	1	5	4.70	5.14
1	1	1	0	5	5.18	5.72
0	0	1	1	9	4.94	5.75
0	1	0	1	3	4.50	2.89
0	1	1	0	2	3.29	2.06
1	0	0	1	2	3.63	1.92
1	0	1	0	3	2.82	2.21
1	1	0	0	14	5.67	13.63
0	0	0	1	8	6.31	8.41
0	0	1	0	3	2.84	3.38
0	1	0	0	8	14.01	8.01
1	0	0	0	4	10.43	4.33
0	0	0	0	34	28.75	33.67

McHugh fitted a latent class model with two classes, obtaining the maximum likelihood estimates using Fishers scoring technique. Everitt (1984a,b) fitted both a two class and a three class model to the data using the EM algorithm. He applied goodness-of-fit tests to both models by calculating the frequencies of the response vectors expected under the model and comparing these with the observed frequencies by the χ^2 statistic. He found that the three class model provided a better fit than the two class model, and this appeared to be mainly due to the improvement in fit of response pattern '1100' between the two models. With the two class model, one class could be interpreted as comprising 'creative' individuals

whilst the other could be interpreted as comprising 'non-creative' individuals. No interpretation was reported for the three class model. Further details can be found in Everitt (1984a,b).

The versatility of latent class analysis as a clustering method was shown by Aitkin, Anderson and Hinde (1981) who fitted two class and three class models to 38 binary variables describing the teaching manner of 468 teachers. They gave a clear interpretation of the latent classes for these models. In the two class model, the latent classes could be interpreted as 'formal' and 'informal' teaching styles. The three class model included both of these types of teaching styles, and the third class shared characteristics of both the 'formal' and 'informal' teaching styles and could be interpreted as a 'mixed' teaching style.

Pickering and Forbes (1984) used this method to study an even larger data set. This 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.

Uebersax and Grove (1990) used latent class methods to analyse diagnostic agreement. In their example, a panel of five diagnosticians rated 859 cases on whether it was appropriate to perform an operation, carotid endartectomy; the ratings were recoded to dichotomies where '+' was a judged indication of treatment and '−' was a judged non-indication of treatment. The data are displayed in Table 1.2.

Table 1.2

Ratings[†] *of treatment appropriateness by physicians.*

Diagnostic Rating					Frequency
+	+	+	+	+	69
+	+	+	+	−	2
+	+	+	−	+	4
+	+	+	−	−	1
+	+	−	+	+	2
+	+	−	+	−	1
+	−	+	+	+	82
+	−	+	+	−	4
+	−	+	−	+	23
+	−	+	−	−	8
+	−	−	+	+	67
+	−	−	+	−	24
+	−	−	−	+	42
+	−	−	−	−	41
−	−	+	+	+	5
−	−	+	−	+	8
−	−	+	−	−	8
−	−	−	+	+	5
−	−	−	+	−	28
−	−	−	−	+	49
−	−	−	−	−	386

[†]Patterns not recorded had 0 frequency.

Uebersax and Grove fitted two, three and four class models to the data and calculated the sensitivity of diagnosis for each rater. They gave an interpretation of 'non indication', 'equivocal indication' and 'valid indication' to the three classes in the three class model. Further references to other studies using latent class methods can be found for example, in Pickering and Forbes (1984), and Uebersax and Grove (1990).

Clustering by latent class analysis does seem to give plausible results in the situations where it has been used, and under some circumstances, it is possible to adjust the model to improve the fit to the data. For instance, suppose that for most i, there is a k such that $\hat{z}_{ik} \approx 1$, where i runs over observations and k runs over groups. In other words, most observations are firmly classified into some group. Then, the local independence assumption can be assessed separately for each group. If the assumption appears to be badly violated, the model can be adjusted either by dropping one of the variables, or by replacing a pair of variables with a new categorical variable with levels indexing the two-way table defined by all pairs of outcomes of the two variables, possibly combining some cells.

The assumption of local independence in latent class analysis makes possible the fitting of highly multivariate mixture models. It also means that the estimated parameters give clear summaries of what may be large and complex sets of data.

1.7 The normal mixture model

Finite mixture models are frequently fitted with the assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a random sample arising from a finite mixture where the component distributions are multivariate normal. That is, $\mathbf{x} \sim N(\boldsymbol{\mu}_k, \Sigma_k)$ with probability π_k for $k = 1, \ldots, K$.

We will only consider the situation where the covariance matrices are not assumed to be equal. Details on the case with equal covariance matrices are given by McLachlan and Basford (1988, section 2.1).

Under the normality assumption, the loglikelihood $L(\boldsymbol{\phi})$, is unbounded, and it is well known that the maximum likelihood estimator of $\boldsymbol{\phi}$ does not exist. For example, consider a mixture of two univariate normal distributions with density function

$$f(\mathbf{x}; \boldsymbol{\phi}) = \pi_1 f_1(\mathbf{x}; \mu_1, \sigma_1^2) + (1 - \pi_1) f_2(\mathbf{x}; \mu_2, \sigma_2^2)$$

Given a random sample of n observations from this density function, the maximum likelihood estimates of the parameters $\boldsymbol{\phi} = (\pi_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ are found by maximising the likelihood function,

$$l(\boldsymbol{\phi}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\phi}).$$

Suppose that $\mu_1 = x_i$ for some i. It can be seen that as σ_1 tends to zero, $l(\boldsymbol{\phi})$ tends to infinity. Thus each observation $\mathbf{x}_i$, gives rise to a singularity in the likelihood function. Hathaway (1985) showed that for mixtures of normal distributions in the univariate case, the sequence of roots corresponding to the largest of the local maxima for each n is consistent and asymptotically normal and efficient. McLachlan and Basford (1988) suggest that it would be surprising if the results given by Hathaway (1985) applying to the univariate case do not carry over to the multivariate case. Further details and references for mixtures of normal distributions are given by McLachlan and Basford (1988), pages 38-39.

We have looked at finite mixtures with normal component distributions, and latent class analysis. Both of these models can be used to fit mixture models to data sets that contain respectively all continuous variables or all categorical variables. However, frequently multivariate data is collected in which both continuous and categorical variables are present. An example of this can be seen in medicine, where continuous measurements are included along with categorical variables indicating

for example, the presence or absence of a certain symptom in a patient, and site of tumour. In the next section, we consider an approach that has been suggested for handling data sets having both categorical and continuous variables. This approach has been motivated by finite mixture models with multivariate normal components.

1.8 Everitt's finite mixture approach

Everitt (1988) proposes that the approach of mixture models with normal components can be adapted for data with both categorical and continuous variables, by assuming that the categorical variables in the data set arise from applying 'thresholds' to underlying unobservable continuous variables. He considers models where the continuous variables, both latent and manifest, have a multivariate normal mixture density.

Suppose that $\mathbf{x} = (x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p+q})'$ is a random vector with p continuous variables and q categorical variables from the finite mixture distribution

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

where $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$, for $k = 1, \ldots, K$, and where the component distributions f_k, are the $N_{p+q}(\boldsymbol{\mu}_k, \Sigma)$ distribution. Note that it is assumed that each group k, has the same covariance matrix.

Now, suppose that $x_{p+1}, \ldots, x_q$ are not directly observable, but are related to a set of observed variables, $z_1, \ldots, z_q$, where

$$z_j = \begin{cases} 1 & \text{if} \quad -\infty = \alpha_{kj1} < x_{p+j} < \alpha_{kj2}, \\ 2 & \text{if} \quad \alpha_{kj2} \leq x_{p+j} < \alpha_{kj3}, \\ \vdots & \qquad\qquad \vdots \\ M_j & \text{if} \quad \alpha_{kjM_j} \leq x_{p+j} < \alpha_{kj(M_j+1)} = \infty, \end{cases}$$

where α_{kjr}, for $k = 1, \ldots, K$, $j = 1, \ldots, q$ and $r = 1, \ldots, M_j$, are the thresholds generating the observed categorical variables from the unobserved continuous variables, $x_{p+1}, \ldots, x_{p+q}$.

By using the conditional density of $x_{p+1}, \ldots, x_q$ given $x_1, \ldots, x_p$, the density function of the observed variables, $x_1, \ldots, x_p$, and $z_1, \ldots, z_q$, can be written in the form

$$h(\mathbf{x}, \mathbf{z}) = \sum_{k=1}^{K} \pi_k N_p(\boldsymbol{\mu}_k, \Sigma) \times \int_{a_1}^{b_1} \cdots \int_{a_q}^{b_q} N_q(\boldsymbol{\mu}_k^{q|p}, \Sigma^{q|p})\, dy_1 \ldots, dy_q$$

where $\boldsymbol{\mu}_k^{q|p}$ and $\Sigma^{q|p}$ are respectively the conditional mean and covariance in group k of $x_{p+1}, \ldots, x_q$ given $x_1, \ldots, x_p$, and the limits of integration (a_i, b_i) for $i = 1, \ldots, q$, equal the threshold values appropriate for the particular values of $z_1, \ldots, z_q$. Further details can be found in Everitt (1988).

Everitt suggested maximising the loglikelihood function for this model using an optimisation algorithm. Everitt illustrated this procedure on four simulated data sets, the largest of which had 100 observations, five variables, two of which were categorical, and required the estimation of 33 parameters. For the data sets analysed, Everitt found that the Simplex method was the most successful in terms of consistent convergence, although convergence could be slow. He suggests the use of the algorithm of Schervish (1984) for categorical variables with more than two levels. Schervish (1984) presents an algorithm for the calculation of multivariate normal probabilities which has error bounds and which includes some time saving devices to speed the calculations. Everitt reports satisfactory performance of the finite mixture approach for the estimation of the models parameters when using the Simplex method.

Such threshold models have been widely used for ordinal data and a brief survey is given in Zhaorong, McGilchrist, and Jorgensen (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].

Everitt and Mérette (1990) demonstrate Everitt's finite mixture method on four simulated data sets, each having three continuous and two categorical variables. They compared the clustering performance of Everitt's finite mixture method with three hierarchical clustering methods where the data had been standardised in a number of ways. They report good performance of the mixture method in comparison to conventional hierarchical methods when applied to simulated data. Everitt and Mérette also applied this method to Fisher's iris data after two of the four variables had been categorised. Further details can be found in Sections 3.3.2 and 3.3.3. They again found that Everitt's finite mixture approach performed better than the hierarchical methods applied.

However, there are some severe practical limitations to the use of this method at present, with the method being basically restricted to data sets that have only one or two categorical variables. Everitt proposes the use of standard optimisation algorithms applied to the loglikelihood function for maximisation. The computation of loglikelihood function requires the numerical evaluation of a q-dimensional integral, where q is the number of categorical variables. Neither Everitt (1988),

nor Everitt and Mérette (1990), consider any examples where the number of categorical variables is greater than two. Clearly this method is not yet ready to take on data sets having very many categorical variables.

In the next section, we describe a model that has been proposed for handling data sets with both categorical and continuous variables.

1.9 The location model

Olkin and Tate (1961) proposed a model that could be used for multivariate data with both categorical and continuous variables. This model, which is known in the literature as the location model, has a distribution function in which the marginal distribution of the discrete variables is multiplied by the conditional distribution of the continuous variables given the values of the discrete. In this model, the discrete variables are arranged in a contingency table form, where the pattern of the discrete variables uniquely determines a cell in the table. The continuous variables are assumed to follow a multivariate normal distribution, where the parameters of the distribution depend on the contingency table cell as defined by the associated discrete variables.

Formally, the model is described as follows. Suppose that a random sample of size n is taken, where each observation has R discrete and p continuous variables. Let $\mathbf{u}' = (u_1, u_2, \ldots, u_R)$ denote the vector of categorical variables, and $\mathbf{v}' = (v_1, v_2, \ldots, v_p)$ denote the vector of continuous variables, measured on each observation. Suppose that the r^{th} categorical variable has m_r categories, then the R categorical variables define an R-way contingency table with $M = \prod_{r=1}^{R} m_r$ cells. Then

(*i*) The n observations have a multinomial distribution over the M cells of the contingency table, and $pr(\text{observation } i \in \text{cell } r) = \lambda_r$ with $\sum_{r=1}^{M} \lambda_r = 1$ for $r = 1, \ldots, M$, and

(*ii*) The conditional distribution of $\mathbf{v}$ given that $\mathbf{u}$ is in cell r, is $N_p(\boldsymbol{\mu}_r, \Sigma)$.

Olkin and Tate (1961) considered various aspects of the location model. In particular, they show that the likelihood function for a sample of n individuals following the location model, factorises into two parts:- one part consists of the multinomial likelihood function with parameters $\lambda_1, \ldots, \lambda_M$, while the other consists of the product of M multivariate normal likelihoods, the r^{th} of which has parameters $\boldsymbol{\mu}_r$ and Σ.

The location model has been used for multivariate data with both categorical and continuous variables by for example, Krzanowski (1983) and (1986) in the discriminant analysis context. Little and Schluchter (1985) describe methods for handling

missing data in the location model. Location models are termed *homogeneous conditional Gaussian* by Lauritzen and Wermuth (1989).

A more general model can be defined by letting the dispersion matrix of the normal distribution vary with the contingency table cell. Krzanowski (1983) included this type of possibility. However, this generalisation leads to a large increase in the number of model parameters. Krzanowski (1988) contends that in practice, the assumption of a common covariance matrix between cells is not unduly restrictive, and that in general, satisfactory results seem to be obtained with the simpler model with the common covariance matrix.

Strictly speaking, the location model in full generality can have several categorical variables, but for programming convenience we have reduced this to one.

We aim to demonstate in this thesis that mixture likelihood methods can be used to cluster highly multivariate data sets where the variables can be either categorical or continuous. We aim to show that our ability to analyse data using the proposed approach can be extended to include data sets where data are missing at random in the sense of Rubin (1976).

In the next chapter, we propose a model that can be used for clustering data sets in which there are both categorical and continuous variables, and in which the number of categorical variables can be greater than two.

Chapter 2

Development of the model

This chapter considers the estimation of the parameters of a finite mixture where the data set consists of both categorical and continuous variables. Section 2.1 discusses the local independence assumption used in the latent class model. Section 2.2 presents the detail for a finite mixture where the component distributions are multivariate normal distributions, and section 2.3 describes the latent class model. Section 2.4 considers the maximum likelihood estimation of the parameters for mixture in which the component distributions are location models. Section 2.5 presents the general approach to multivariate mixture models for multivariate observations on both categorical and continuous variables.

2.1 The local independence assumption

Two major difficulties frustrate the wider application of multivariate normal mixture models. Firstly, they are not easily adapted to cope with discrete data, and most real clustering problems involve data sets containing both continuous and discrete variables. Secondly, they lead to models with large numbers of parameters. For example, in fitting a mixture of two groups to a data set containing four variables, we need to estimate 19 parameters with the assumption of equal covariance matrices for the component distributions; an additional ten parameters need to be estimated if the covariance matrices are not assumed to be equal for the two groups. Highly parameterized models can lead to computational difficulties such as multiple roots of the likelihood equation.

When background information is available from subject-area theory or from previous statistical analysis of similar data sets, it may be possible to specify component distribution functions that are not highly parameterized and that are believed to represent the true shape of the components. However, this thesis is concerned with exploratory data analyses, where little may be known *a priori* about the structure of the data. What is needed in these situations, is a flexible, but not overly flexible, family of multivariate distributions that can be used as a 'default' for the component distributions, in the absence of knowledge that would justify a more detailed specification. Inspiration is drawn from the latent class model that has been used with some success to cluster categorical data.

The form of the component probability functions adopted in the latent class model implies that conditional on membership in a group, the variables are independent.

This *local independence* property is a very strong condition that often causes concern to those considering the use of latent class analysis, because it is not often found in observable sub populations, or even in groups formed by clustering using latent class analysis itself. Should we contemplate fitting models that we know to be oversimplified? Some insight into this problem can be obtained from results of Scott and Symons (1971) for clustering continuous data. They show that certain criteria used for clustering can be derived from likelihood ratio or Bayesian classification methods for multivariate normal mixture models. This result can be looked at in two ways: it can be seen as suggesting that when we choose a clustering criterion in a non hierarchical cluster analysis we are implicitly choosing a certain mixture model for the data, or alternatively that all we are really doing when we choose a nominal model for the component distributions is to indirectly define a new clustering criterion. We would like the clusters defined by the mixture model not to be highly sensitive to the details in the model. Note that as discussed in Chapter 1, the results of Scott and Symon do not directly apply to the maximum likelihood estimation in finite mixture models, as their maximisation is over all possible assignments of observations to populations.

In any event clustering by latent class analysis does seem to give plausible results in many situations where it has been used. Under some circumstances it is even possible to adjust the model to improve the fit to the data. Suppose that most observations are firmly classified to some group k, not necessarily the same group. That is, for most observations i, there is a group k such that the estimated posterior probability of i belonging to that group k is close to one. The assumption of local independence can then be assessed separately for each group. If the assumption appears to be badly violated, the model can be adjusted either by dropping one of the variables, or by replacing a pair of variables with a new categorical variable with levels indexing the two-way table defined by all pairs of outcomes of the two variables, possibly combining some cells.

The assumption of local independence in latent class analysis makes possible the fitting of highly multivariate mixture models. It also means that the estimated parameters give clear summaries of what may be large and complex sets of data. One way of generalising the latent class model to data sets combining categorical and continuous data would be to adjoin to the products of discrete distributions in the latent class model, additional univariate normal factors for each continuous variable. The fitting of models of this kind are looked at in Section 2.4, but these models will not be sufficiently general for our purposes.

For example, in clinical data, frequently both systolic and diastolic blood pressure measurements are available. It is clear that these variables are correlated,

even within diagnostic groups. We will want to assume component distributions that at least reflect likely features of such 'real' sub-populations. Clearly, then, a covariance parameter for the two blood pressures should be in the model.

In the next section, we present the detail of a finite mixture where the component distributions are multivariate normal distributions.

2.2 Mixtures of multivariate normals

Suppose that p attributes are measured on n individuals. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the observed values of a random sample from a mixture of K underlying populations in unknown proportions $\pi_1, \ldots, \pi_K$. Let the density of $\mathbf{x}_i$ in the k^{th} group be $f_k(\mathbf{x}; \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ is the parameter vector for group k, and let $\boldsymbol{\phi} = (\boldsymbol{\theta}', \boldsymbol{\pi}')'$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)'$. The density of $\mathbf{x}$ can be written as

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

where $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$, for $k = 1, \ldots, K$.

The likelihood function for the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is given by

$$l = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right\}.$$

Under the normality assumption, the component densities, $f_k(\mathbf{x}; \boldsymbol{\theta}_k)$, are the $N_p(\boldsymbol{\mu}_k, \Sigma_k)$ densities. As described by McLachlan and Basford (1988) and in Chapter 1, the likelihood function for normal densities with unequal covariance matrices, is unbounded, since each data point gives rise to a singularity on the boundary of the parameter space. However, it is known that there is a sequence of roots of the likelihood equation that is consistent and asymptotically efficient for $\boldsymbol{\phi}$, and with probability tending to one, these roots correspond to local maxima of the likelihood function. Let $\hat{\boldsymbol{\phi}}$ denote the maximum likelihood estimate of $\boldsymbol{\phi}$. Then each observation, $\mathbf{x}_i$, can be allocated to a group on the basis of the estimated posterior probabilites. The posterior probability that observation i belongs to group k is given by

$$\begin{aligned} \tau_k(\mathbf{x}_i; \hat{\boldsymbol{\phi}}) &= pr(\text{observation } i \in \text{ group } k \mid \mathbf{x}_i; \hat{\boldsymbol{\phi}}) \\ &= \frac{\hat{\pi}_k f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k)}{\sum_{k=1}^{K} \hat{\pi}_k f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k)} \end{aligned}$$

for $k = 1, \ldots, K$; and $\mathbf{x}_i$ is assigned to group k if

$$\tau_k(\mathbf{x}_i; \hat{\boldsymbol{\phi}}) > \tau_{k'}(\mathbf{x}_i; \hat{\boldsymbol{\phi}}) \text{ for } k = 1, \ldots, K; k \neq k'.$$

For convenience, $\tau_k(\mathbf{x}_i; \hat{\boldsymbol{\phi}})$ will be denoted as $\hat{z}_{ik}$.

The EM algorithm is applied to the finite mixture model by viewing the data as incomplete. As described in Chapter 1, the EM algorithm works by the conceptual adjoining of 'missing data' onto the observed data to form the 'complete data', for which maximum likelihood estimation is simple. In the case of mixtures of distributions, the 'missing' data are the unobserved indicators of group membership. Let the vector of indicator variables, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$, be defined by

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

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 hypothetical 'complete data', then, consists of the $n \times p$ array of observed data together with the conceptual $n \times K$ array $\{z_{ik}\}$, of class membership indicators.

The loglikelihood for the hypothetical 'complete data' is

$$\begin{aligned} L_C(\boldsymbol{\phi}) &= \log\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\left[\pi_k^{z_{ik}}\left\{f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\}^{z_{ik}}\right]\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{z_{ik}\log\pi_k + z_{ik}\log f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\pi_k - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\Big\{(\mathbf{x}_i-\boldsymbol{\mu}_k)'\,\Sigma_k^{-1}\,(\mathbf{x}_i-\boldsymbol{\mu}_k) \\ &\qquad + \log\left((2\pi)^p\,|\Sigma_k|\right)\Big\} \qquad (2.1) \end{aligned}$$

where log denotes the natural logarithm.

It can be seen that the term in the curly brackets, {}, in (2.1) is in the form of a quadratic in x_{ij}. The 'complete' data sufficient statistics for the model are

(*i*) $\sum_{i=1}^{n} z_{ik}$, for each group k;

(*ii*) $\sum_{i=1}^{n} z_{ik}x_{ij}$ for each group k, and each variable x_{ij};

(*iii*) $\sum_{i=1}^{n} z_{ik}x_{ij}x_{ij'}$ for each group k, and each pair of variables, x_{ij} and $x_{ij'}$.

The E step of the EM algorithm requires the calculation of

$$Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)}) = E\{L_C(\boldsymbol{\phi}) \mid \mathbf{x}; \boldsymbol{\phi}^{(t)}\},$$

the expectation of the complete data loglikelihood, conditional on the observed data and the current value of the parameters. We calculate $Q(\phi, \phi^{(t)})$ from $L_C(\phi)$ by replacing z_{ik} with

$$\begin{aligned}\hat{z}_{ik} &= E(z_{ik} \mid \mathbf{x}_i\,; \phi^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_i\,; \theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i\,; \theta_k^{(t)})}\end{aligned}$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that observation i belongs to group k.

In the M step of the algorithm, $\phi^{(t+1)}$ is chosen to be a value of ϕ which maximises $Q(\phi, \phi^{(t+1)})$ with respect to ϕ. The components of ϕ for this model are given by

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ik}$$

$$\hat{\mu}_{kj} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik} x_{ij}$$

$$\hat{\Sigma}_{kjj'} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik} \left[(x_{ij} - \hat{\mu}_{kj})(x_{ij'} - \hat{\mu}_{kj'})\right]$$

for $j, j' = 1, \ldots, p$ and $k = 1, \ldots, K$.

The EM algorithm alternates between the two calculations, the E step and the M step until convergence. Further details on mixtures of multivariate normal distributions are given by McLachlan and Basford (1988).

In the next section, we present the detail for the latent class model, expressed as a finite mixture.

2.3 The latent class model

Suppose that the population is made up of K groups or latent classes, in proportions $\pi_1, \ldots, \pi_K$, where $\pi_k \geq 0$ and $\sum \pi_k = 1$. Let $\mathbf{x} = (x_1, \ldots, x_p)'$ be the response vector observed on each observation, where the jth variable can have levels numbered from 1 to M_j. Let λ_{kjm} be the probability that variable j takes level m in group k. Then, if the ith observation $\mathbf{x}_i$, happens to come from group k, its probability function is given by

$$f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) = \prod_{j=1}^{p} \lambda_{kjx_{ij}}$$

where $\boldsymbol{\theta}_k$, the vector of unknown parameters of the distribution of the responses in the kth sub-population, are in this case, the $\{\lambda_{kjm}\}$.

The overall probability function is a mixture of these conditional probability functions:

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

The parameter vector $\boldsymbol{\phi}$ is made up of the π_k and the λ_{kjm} as k, j, and m take on all allowable values. Note that we have overparameterized here, as the π_k summed over k and the λ_{kjm} summed over m for any fixed j, k will total 1.

Suppose that for $j = 1, \ldots, p$, we define an indicator variable

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

for $m = 1, \ldots, M_j$. Then, the density function for observation i in group k, can be written as

$$f_k(\mathbf{x}_i;\theta_k) = \prod_{j=1}^{p} \prod_{m=1}^{M_j} \lambda_{kjm}^{\delta_{ijm}}.$$

The EM algorithm is applied to the finite mixture model where the indicator variables $\mathbf{z}_1, \ldots, \mathbf{z}_n$, are as described in section 2.2. The hypothetical 'complete data' consists of the $n \times p$ array of observed data together with the conceptual $n \times K$ array $\{z_{ik}\}$, of class membership indicators.

The likelihood for the complete data set can be written as

$$\begin{aligned} \ell_C(\boldsymbol{\phi}) &= \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}} \left[f_k(\mathbf{x};\theta_k)\right]^{z_{ik}} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}} \left\{ \prod_{j=1}^{p} \prod_{m=1}^{M_j} \lambda_{kjm}^{\delta_{ijm}} \right\}^{z_{ik}} \end{aligned}$$

The loglikelihood for the hypothetical complete data is

$$L_C(\boldsymbol{\phi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left(z_{ik} \left\{ \log \pi_k + \sum_{j=1}^{p} \sum_{m=1}^{M_j} \delta_{ijm} \log(\lambda_{kjm}) \right\} \right).$$

From $L_C(\boldsymbol{\phi})$, we see that the complete data sufficient statistics for the model are,

(*i*) $\sum_{i=1}^{n} z_{ik}$ for each group k, and

(*ii*) $\sum_{i=1}^{n} z_{ik}\delta_{ijm}$ for each group k, each categorical variable x_j, and each value m of x_j.

The E step of the EM algorithm requires the calculation of

$$Q(\phi, \phi^{(t)}) = E\{L_C(\phi) \mid \mathbf{x}; \phi^{(t)}\},$$

the expectation of the complete data loglikelihood, 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 $Q(\phi, \phi^{(t)})$ by replacing z_{ik} with

$$\begin{aligned}\hat{z}_{ik} &= E(z_{ik} \mid \mathbf{x}_i\,; \phi^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_i\,; \theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i\,; \theta_k^{(t)})}\end{aligned}$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that observation i belongs to group k.

At the M step, $\phi^{(t+1)}$ is chosen to be the value of ϕ that maximises $Q(\phi, \phi^{(t)})$ with respect to ϕ.

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ik}$$

$$\hat{\lambda}_{kjm} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik} \delta_{ijm}$$

for $k = 1, \ldots, K$, $j = 1, \ldots, p$ and $m = 1, \ldots, M_j$.

The EM algorithm alternates between the E step and the M step until convergence.

In Section 1.8, we described the finite mixture approach suggested by Everitt (1988) for clustering data sets having both categorical and continuous variables. As was pointed out, this approach was basically restricted to data sets having only one or two categorical variables. In the next section, we consider another approach for the maximum likelihood estimation of the parameters for a finite mixture where the data set consists of both categorical and continuous variables. Specifically, we fit a mixture in which the component distributions are location models. For our model, we shall assume that the location model has only one categorical variable. Strictly speaking, the location model in full generality can have several categorical variables, but for programming convenience we have reduced this to one.

2.4 Mixtures of location models

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$, be a random sample from the finite mixture distribution

$$f(\mathbf{x}; \phi) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}; \theta_k)$$

where $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$, for $k = 1, \ldots, K$.

Suppose that $p+1$ attributes are measured on the n individuals, where the vector of variables, $\mathbf{x} = (x_1, \ldots, x_{p+1})'$, has p continuous variables, and one discrete variable. The component distributions, $f_k(\mathbf{x}; \boldsymbol{\theta}_k)$, have the following form. The discrete variable takes the values $1, \ldots, M$ with probabilities $\lambda_{k1}, \ldots, \lambda_{kM}$, where $\sum_{m=1}^{M} \lambda_{km} = 1$. Conditional on the discrete variable taking value m, the p continuous variables have the multivariate normal distribution, $N_p(\boldsymbol{\nu}_{km}, \Xi_k)$. This is the location model of Olkin and Tate (1961), and described in Section 1.9.

To distinguish between the categorical and continuous variables, in this section, u will be used to denote the discrete variable and $\mathbf{v}$ will be used to denote the $p \times 1$ vector of continuous variables. Thus, the observation vector for observation i, $\mathbf{x}_i$ takes the form $(u_i, \mathbf{v}_i)$.

The EM algorithm is applied to the finite mixture model, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ for $i = 1, \ldots, n$, are the indicator variables as described in Section 2.2. The hypothetical 'complete data' consists of the observed data together with the conceptual $n \times K$ array $\{z_{ik}\}$ of class membership indicators.

The loglikelihood for the hypothetical 'complete data' is

$$
\begin{aligned}
L_C(\boldsymbol{\phi}) &= \log \left(\prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k^{z_{ik}} \left\{ f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right\}^{z_{ik}} \right] \right) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ z_{ik} \log \pi_k + z_{ik} \log f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right\} \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \left[z_{ik} \log \pi_k + z_{ik} \left\{ \sum_{m=1}^{M} w_{im} \log \lambda_{km} + h(\Xi_k) \right. \right. \\
&\qquad \left. \left. -\frac{1}{2} \left(\mathbf{v}_i - \sum_{m=1}^{M} w_{im} \boldsymbol{\nu}_{km} \right)' \Xi_k^{-1} \left(\mathbf{v}_i - \sum_{m=1}^{M} w_{im} \boldsymbol{\nu}_{km} \right) \right\} \right]
\end{aligned}
$$

where $h(\Xi_k) = \frac{1}{2} np \log(2\pi) + \log(|\Xi_k|)$, and $w_{im} = \begin{cases} 1 & \text{if } u_i = m, \\ 0 & \text{otherwise.} \end{cases}$

By inspection of the preceding expression for $L_C(\boldsymbol{\phi})$, the complete data sufficient statistics for the model are

(*i*) $\sum_{i=1}^{n} z_{ik}$, for each group k;

(*ii*) $\sum_{i=1}^{n} z_{ik} w_{im} v_{ij}$ for each class k, each continuous variable v_j and each value m of the categorical variable u;

(*iii*) $\sum_{i=1}^{n} z_{ik} v_{ij} v_{ij'}$, for each class k, each pair of continuous variables v_j and $v_{j'}$, for $j \leq j'$.

The EM algorithm alternates between the two calculations, the E step and the M step. At the t^{th} iteration, let $\theta_k^{(t)} = (\lambda_k^{(t)}, \nu_{mk}^{(t)}, \Xi_k^{(t)}\,;\, m = 1, \ldots, M)$ denote the current estimates of the parameters for group k. The E step of the algorithm requires the calculation of $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)})$, the expectation of the complete data loglikelihood, conditional on the observed data and the current value of the parameters.

We calculate $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)})$ by replacing z_{ik} with

$$\begin{aligned} \hat{z}_{ik} &= E(z_{ik} \mid \mathbf{x}_i\,;\boldsymbol{\phi}^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_i\,;\theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i\,;\theta_k^{(t)})} \end{aligned}$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that observation i belongs to group k.

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

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ik}$$

$$\hat{\lambda}_{km} = \frac{1}{n\hat{\pi}_k} \sum_{i, u_i = m} \hat{z}_{ik}$$

$$\begin{aligned} \hat{\nu}_{kjm} &= \frac{1}{n\hat{\pi}_k} E\Big(\sum_{i=1}^{n} z_{ik} w_{im} v_{ij} \mid \mathbf{x};\, \theta_k\Big) \\ &= \frac{1}{n\hat{\pi}_k} \sum_{i, u_i = m} \hat{z}_{ik} v_{ij} \end{aligned}$$

$$\begin{aligned} \hat{\Xi}_{kjj'} &= \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \sum_{m=1}^{M} \hat{z}_{ik} w_{im} (v_{ij} - \hat{\nu}_{kjm})(v_{ij'} - \hat{\nu}_{kj'm}) \\ &= \frac{1}{n\hat{\pi}_k} \sum_{i, u_i = m} \hat{z}_{ik} (v_{ij} - \hat{\nu}_{kjm})(v_{ij'} - \hat{\nu}_{kj'm}) \end{aligned}$$

for $k = 1, \ldots, K$; $m = 1, \ldots, M$; and $j, j' = 1, \ldots, p$.

The EM algorithm alternates between the E step and the M step until convergence.

In the next section, we describe a general class of multivariate mixture models for multivariate observations on both categorical and continuous variables. This approach is based on a model suggested by latent class analysis, and generalises both latent class and multivariate normal models.

2.5 A general approach to multivariate mixture models.

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

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

which is a finite mixture of the K component distributions f_k and where $\pi_k \geq 0$ and $\sum \pi_k = 1$.

The distributions $f_k(\mathbf{x};\boldsymbol{\theta}_k)$ must be kept simple in structure for two reasons. Firstly we would like the model to give us an understandable decomposition of the data that aids us in visualising the data. Secondly the f_k must be restricted if we are to be able to have any hope of identifying the parameters π_k, if this were not so then corresponding to any decomposition $f = \sum \pi_k f_k$ we could consider another decomposition where, for example, $f = f_1$, $\pi_1 = 1$ and $\pi_2 = \ldots = \pi_K = 0$.

The simple structure that we choose is based on local independence. We suppose that the vector of variables $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p)'$ has been partitioned so that

$$\mathbf{x} = (\breve{\mathbf{x}}_1 | \ldots | \breve{\mathbf{x}}_l | \ldots | \breve{\mathbf{x}}_L)'$$

then we will consider component distributions of the form

$$f_k(x) = \prod_{l=1}^{L} f_{kl}(\breve{\mathbf{x}}_l).$$

We will refer to the subvector of variables $\breve{\mathbf{x}}_l$ as the lth *partition cell*, to avoid confusion with classes or groups of observations or cells in tables.

The form of local independence that we are assuming is that within each of the K sub populations, the variables within the partition cell $\breve{\mathbf{x}}_l$, are independent of the variables in partition cell $\breve{\mathbf{x}}_{l'}$, for $1 \leq l < l' \leq L$. The functions f_{kl} form the 'atoms' out of which our model is built by crossing and mixing. The present work uses the following distributions for the $\breve{\mathbf{x}}_{kl}$, but it should be stressed that to a considerable degree the choice is arbitrary.

(a) *Discrete Distribution*

Where $\breve{\mathbf{x}}_l = \{x_j\}$ is a 1-dimensional discrete random variable taking values $1, \ldots, M_l$ with probabilities $\lambda_{kl1}, \ldots, \lambda_{klM_l}$. We will denote this distribution by $D(\lambda_{kl1}, \ldots, \lambda_{klM_l})$. If all f_{kl} are of this form, then f is a latent class model.

(b) *Multivariate Normal*

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

(c) *Location Model*

Where $\breve{\mathbf{x}}_l$ is a $1+p_l$ dimensional vector of random variables with one discrete variable, u_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_j, the p_l continuous random variables have the multivariate normal distribution $N_{p_l}(\boldsymbol{\nu}_{mkl}, \Xi_{kl})$.

The model for the ith observation can thus be written as

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

where $\boldsymbol{\theta}_{kl}$ consists of the parameters of the distribution f_{kl} as described above.

Note that in each of the K classes or subpopulations, the vector random variable $\breve{\mathbf{x}}_l$, of the lth partition cell has the same type, either (a) or (b) or (c), but the parameters may vary from group to group.

It can be seen that in fitting the model to a particular data set, there is considerable discretion in how the partition is formed. In general, the larger the dimensions of the $\breve{\mathbf{x}}_l$, the more covariance parameters must be added to the model, and the poorer the stability of the parameter estimates. On the other hand too few covariances in the model will result in a poor fit, which may or may not have consequences for the cluster assignments. A reasonable model selection strategy appears to be to begin with the model with complete local independence and fit it for a few values of K, the number of classes. Then variables with strong within-cluster associations can be grouped together in a partition cell for the next series of fits, and so on.

It is convenient to use the notation u_l for the discrete variable when the lth partition cell is of type (a) or (c), and $\mathbf{v}_l$ for the $p_l \times 1$ vector of continuous variables when the lth partition cell is of type (b) or (c). So in cases (a), (b) and (c), $\breve{\mathbf{x}}_l$ takes the forms u_l, $\mathbf{v}_l$ and $(u_l, \mathbf{v}_l)$ respectively.

The model, as described above, 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, Laird and Rubin (1977).

Let the indicator vectors $\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_n$ be as described in section 2.2. The

complete-data specification treats the $\mathbf{z}_i$ as known leading to the loglikelihood

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

where

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

Maximising the loglikelihood $L_C(\phi)$, for the complete data is equivalent to maximising $l_k(\boldsymbol{\theta}_k)$ separately for each partition cell. For the details of the complete data sufficient statistics for partition cells with

(a) *The discrete distribution*

Refer to section 2.3.

(b) *The Multivariate Normal Distribution*

Refer to section 2.2.

(c) *The Location Model*

Refer to section 2.4.

The sufficient statistics for each partition cell can now be amalgamated such that the complete data sufficient statistics for the model are

(*i*) $\sum_{i=1}^{n} z_{ik}$, for each group k;

(*ii*) $\sum_{i=1}^{n} z_{ik}\delta_{ilm}$, for each class k, each categorical variable u_l, and each value m_l of u_l, where $\delta_{ilm} = \begin{cases} 1 & \text{if } x_{il} = m_l, \\ 0 & \text{otherwise.} \end{cases}$

(*iii*) Multivariate Normal partition cells

a. $\sum_{i=1}^{n} z_{ik}v_{ij}$ and $\sum_{i=1}^{n} z_{ik}v_{ij}^2$, for each group k, and each continuous variable v_j belonging to a multivariate normal partition cell.

b. $\sum_{i=1}^{n} z_{ik}v_{ij}v_{ij'}$, for each group k, and each pair of continuous variables, v_j and $v_{j'}$, $j < j'$, belonging to the same multivariate normal partition cell.

(*iv*) Location Model partition cells

a. $\sum_{i=1}^{n} z_{ik}w_{ilm}v_{ij}$ for each group k, each continuous variable v_j and each value m

of the categorical variable u_l, belonging to the same location model partition cell;

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

b. $\sum_{i=1}^{n} z_{ik} v_{ij} v_{ij'}$, for each class k, each pair of continuous variables v_j and $v_{j'}$, $j \leq j'$, belonging to the same location model partition cell.

The EM iteration alternates between two calculations, the E step and the M step. Beginning at a current value for $\boldsymbol{\phi}$, say $\boldsymbol{\phi}^{(t)}$, the vector of all unknown parameters, the E step requires the calculation of

$$Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)}) = E\{L_C(\boldsymbol{\phi}) \mid \mathbf{x}; \boldsymbol{\phi}^{(t)}\},$$

the expectation of the complete data loglikelihood, conditional on the observed data and the current value of the parameters. We calculate $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)})$ by replacing z_{ik} with

$$\begin{aligned} \hat{z}_{ik} &= E(z_{ik} \mid \mathbf{x}_i; \boldsymbol{\phi}^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_i, \theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i, \theta_k^{(t)})} \end{aligned}$$

That is, z_{ik} is replaced by z_{ik}, the estimate of the posterior probability that observation i belongs to group k.

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

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ik}$$

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

$$\hat{\boldsymbol{\mu}}_{kl} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik} \mathbf{v}_{il}$$

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

$$\hat{\boldsymbol{\nu}}_{klm} = \frac{1}{n\hat{\pi}_k} \sum_{i, u_{il}=m} \hat{z}_{ik} \mathbf{v}_{ijm}$$

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

for $k = 1, \ldots, K$, $m = 1, \ldots, M_l$ 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 partition cell.

A program has been written in Fortran to implement the approach of multivariate mixture models. The program requires:- (i) the data in the form of an $n \times p$ matrix of observations by variables, and (ii) an input file, giving details of the number of observations in the dataset, number of variables, type of variables, *et cetera.* Further details of the format of the input file are given in Appendix 1. A program has been written to create the input file in the required format.

The iteration may be started either from an initial classification or from an initial set of parameter estimates. If the number of parameters is quite large, it is usually more convenient to begin with a classification. The current version of the program uses a convergence criterion to cease iterating when the difference in loglikelihoods at iteration t and iteration $t - 10$ is 0.0000001. This can be altered if needed. We have used this convergence criterion as the scaled distance between the parameters is similar to a difference between the loglikelihoods. McLachlan and Basford (1988) in program KMM compare the difference in loglikelihoods at iteration t and iteration $t - 10$ with $0.0001\times$ (loglikelihood at iteration $t - 10$).

The model, as described, is a mixture of K distributions, but it is interesting to note that it can be described in the language of graphical models used by Lauritzen and Wermuth (1989): if we draw a graph with vertices for each variable, and an extra vertex for the latent variable giving the class assignment, then variables in the same partition cell form a *clique* (maximal complete subgraph), all variables are connected to the latent variable, and variables in different partition cells are connected to each other only by a path through the latent variable.

For fully categorical datasets, the method of Latent Class Analysis has become a popular method of discovering underlying cluster structure. The class of models described in this section, forms a natural extension of the Latent Class model to data sets containing both categorical and continuous variables. Like Latent Class models, this model makes free use of local independence to reduce the number of parameters in the model and to lead to descriptions of the clusters that can be easily understood. Provision is made, however, for the cautious introduction of within-cluster covariances.

Because the EM algorithm is used to fit the models it is feasible to fit them to datasets with many variables and observations, so that for many applications, fitting these models becomes an alternative to conventional cluster analysis algorithms. In Chapter 4, this approach is extended to cope in situations where data

are missing at random.

In the next chapter, we demonstrate the approach of multivariate finite mixture models by clustering three data sets.

Chapter 3

Using the Multivariate Mixture Model

3.1 Introduction

The use of multivariate mixture models for clustering is not new but has usually been applied to data sets of fairly low dimension. In this chapter, we demonstrate that mixture model likelihood methods can be used to usefully cluster multivariate data sets where the variables can be either categorical or continuous. In section 3.2, we illustrate the scope of this method by considering the clustering of cases on the basis of pre-trial variables alone for the Prostate Cancer clinical trial data of Byar and Green (1980) reproduced in Andrews and Herzberg (1985, pp 261–247). In section 3.3, we consider the clustering of Fishers (1936) Iris data, and in section 3.4, we look at the clustering of observations on the basis of the results of a statistics examination.

3.2 The Cancer data

This data was obtained from a randomized clinical trial comparing four treatments for 506 patients with prostatic cancer. These patients had been grouped by physicians using clinical criteria into Stage 3 and Stage 4 of the disease. Patients were classified as having Stage 3 prostatic cancer when the cancer was contained locally and there was no clinical evidence that the disease had spread elsewhere, whilst they were classified as having Stage 4 prostatic cancer when there was evidence that the disease had spread to other parts of the body. This evidence could have been obtained from elevated acid phosphatase levels, x-rays, or both of these. Byar and Green (1980) originally considered this data in their investigation to determine whether there may be an optimal treatment for each patient based on his individual characteristics. They approached this question by looking for treatment-covariate interactions in an exponential survival model. They found that younger patients with high grade tumours should have been treated with estrogens, whereas older patients with low grade tumours were likely to be harmed by estrogen treatment.

In the analysis that follows, we shall use the multivariate mixture models approach to clustering this data set, not because this is necessarily the best way to answer the question studied by Byar and Green, but to demonstrate that the maximum likelihood fitting of multivariate mixture models is a viable approach to clustering

data sets of this size and complexity. The clusters that are obtained from using the approach of multivariate mixture models will be compared with the grouping that is available from the clinical classification. The information about the clinical classification will be excluded from the analysis, and the pre-trial variables will be used to cluster the cases. The clusters formed from using this approach, will be examined to see if they have any reality in terms of the underlying disease status of the patient. The grouping structure will also be investigated by considering the survival status and the trial outcome of patients in different clusters.

There are twelve pre-trial covariates 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.

Table 3.1

Pretreatment covariates

Covariate	*Abbreviation*	*Number of Levels (if categorical)*
Age	Age	
Weight	Wt	
Performance rating	PF	4
Cardiovascular disease history	HX	2
Systolic Blood pressure	SBP	
Diastolic blood pressure	DBP	
Electrocardiogram code	EKG	7
Serum haemoglobin	HG	
Size of primary tumour	SZ	
Index of tumour stage and histolic grade	SG	
Serum prostatic acid phosphatase	AP	
Bone metastases	BM	2

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. A square root transformation was used for SZ, and a logarithmic transformation was used for AP to achieve approximate normality. (As for correlation, skewness over the whole data set does not necessarily mean skewness within clusters but when clusters were formed within-cluster skewness was observed for these variables.) 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 chapter

4, we extend the multivariate mixture model to cope with data missing at random in the sense of Little and Rubin (1987), and in chapter 5, we describe the analyses for the pretreatment covariates of the cancer data set with missing values included.

3.2.1 Two-component models.

Because the data has already been classified by physicians using clinical criteria, into two groups, Stage 3 (no clinical evidence of metastasis) and Stage 4 (clinical evidence of metastasis), we shall fit a mixture of two components to see the extent to which the program can rediscover these stages. The model does not specify a common covariance structure for the covariates. We regard the data as a random sample from the distribution

$$f(\mathbf{x};\boldsymbol{\phi}) = \sum_{k=1}^{2} \pi_k f_k(\mathbf{x};\boldsymbol{\theta}_k)$$

where $\sum_{k=1}^{2} \pi_k = 1$, and $\pi_k \geq 0$, $k = 1,2$. Under the model with complete local independence for two clusters, the component distributions will be of 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, partition l, and $k = 1,2$. We see that $f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl})$ is $N(\mu_{kl},\sigma_{kl}^2)$ for each of the eight continuous variables, and $D(\lambda_{kl1},\ldots,\lambda_{klm_l})$[1] for each of the four categorical variables.

This model was fitted iteratively using the EM algorithm with the initial estimates of the group parameters being based on those resulting from the clinical classification. An observation $\mathbf{x}_i$ is assigned to the population to which it has the highest estimated posterior probability of belonging; that is, we assign $\mathbf{x}_i$ to group k if

$$\tau_k(\mathbf{x}_i;\hat{\boldsymbol{\phi}}) \geq \tau_{k'}(\mathbf{x}_i;\hat{\boldsymbol{\phi}}) \qquad \text{for } k' = 1,2;$$

where

$$\tau_k(\mathbf{x}_i;\boldsymbol{\phi}) = \text{pr}\left(i^{th}\text{observation} \in k \mid \mathbf{x}_i;\boldsymbol{\phi}\right) = \pi_k f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)/\{\sum_{k=1}^{2} \pi_k f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\}.$$

We will use $\hat{z}_{ik}$ to denote $\tau_k(\mathbf{x}_i;\hat{\boldsymbol{\phi}})$. The following general notation will be used for the models fitted. 'Model i, n' denotes the model that has partitioning i, and n groups. For the cancer data set, the number of groups fitted to the model will be omitted from the notation when two groups are fitted. Hence 'Model i' refers to the model that has partitioning i and two groups, and is equivalent to 'Model i, 2'. Using this notation, the model with complete local independence will be referred to as Model 1.

[1] refer to notation in Chapter 2.

Table 3.2

Agreements and differences between the clinical and model classifications for the model with complete local independence, Model 1.

		Model Classification	
	Class	**Group 1**	**Group 2**
Clinical Classification	**Stage 3**	252	21
	Stage 4	20	182

It was found for Model 1, (see Table 3.2), that the clinical classification and the 'statistical diagnosis' are different for 41 observations out of the 475. If we take the posterior probability $\hat{z}_{ik}$ to be at least 0.95 to classify an observation as being 'definitely assigned' to a group, we find that 15 of the 41 are definitely assigned to a different group than the one corresponding to the clinical classification. Eight observations have a greatest posterior probability lying between 0.5 and 0.7. Another comparison between the clinical classification into stages 3 and 4 and the Model 1 fit can be obtained by comparing the estimated parameters under Model 1 with their counterparts under the clinical classification. This is done in Table 3.3 and it may be seen that the agreement is quite close. The estimated proportions in the two classes are $\hat{\pi}_1 = .5637$, and $\hat{\pi}_2 = 0.4363$ leading to expected numbers in each class of 267.8 and 207.2. This is quite close to the 273/202 split in the clinical classification.

As the likelihood equation for mixture models usually has multiple roots, the EM Algorithm should be applied over a wide range of starting values in any search of local maxima. In order to search for other maxima, and to dispel any suspicion that the estimated parameters are close to the statistics for the clinical classification merely because these where used as starting values, the algorithm was run again 10 more times from initial classifications generated by randomly splitting the patients into two groups. Three solutions of the likelihood equation were found for Model 1. From ten starting values, seven converged to a solution with a loglikelihood of -11386.265, the same solution that was found using the parameters based on the clinical classification. Two iterations converged to a solution with a loglikelihood of -11476.051, and one iteration converged to a solution with a loglikelihood of -11392.972.

Model 1 is relatively easy to fit because of the small number of parameters. It is also easy to comprehend because the dependence between variables is totally explained by the cluster structure. Once this model has been fitted, we can seek ways of improving the fit by adding more covariance parameters.

Table 3.3

Summary statistics of the 12 pretreatment variables according to the clinical classification and the Model 1 estimates.

(i) Continuous variables.

	Clinical Classification			Model Classification		
Variable	**Stage**	**MEAN**	**STDEV**	**Group**	**MEAN**	**STDEV**
Age	3	71.9	6.490	1	71.6	6.658
	4	71.1	7.454	2	71.5	7.229
Wt	3	100.1	13.090	1	100.5	13.075
	4	97.5	13.548	2	97.2	13.419
SBP	3	14.5	2.569	1	14.5	2.557
	4	14.3	2.232	2	14.2	2.232
DBP	3	8.2	1.542	1	8.3	1.543
	4	8.1	1.357	2	8.0	1.336
HG	3	137.2	18.520	1	138.1	17.813
	4	130.1	19.820	2	129.1	20.093
sqrt(SZ)	3	3.087	1.377	1	2.896	1.201
	4	3.915	1.687	2	4.140	1.702
SG	3	9.06	1.348	1	8.9	1.155
	4	12.0	1.504	2	12.1	1.425
log(AP)	3	1.622	0.548	1	1.647	0.500
	4	4.014	1.612	2	3.921	1.697

(ii) Categorical variables.

Level probabilities from the Clinical classification.

Variable	Stage	Level probabilities						
PF	3	0.934	0.055	0.011	0.0			
	4	0.856	0.084	0.050	0.010			
HX	3	0.524	0.476					
	4	0.619	0.381					
EKG	3	0.348	0.051	0.077	0.059	0.297	0.165	0.004
	4	0.327	0.045	0.144	0.045	0.317	0.124	0.00
BM	3	0.996	0.004					
	4	0.624	0.376					

Level probabilities from the Model Classification using Model 1.

Variable	Group	Level probabilities						
PF	1	0.940	0.052	0.008	0.0			
	2	0.850	0.087	0.053	0.010			
HX	1	0.493	0.507					
	2	0.656	0.344					
EKG	1	0.332	0.046	0.075	0.061	0.311	0.171	0.004
	2	0.348	0.052	0.144	0.042	0.297	0.117	0.00
BM	1	0.992	0.008					
	2	0.639	0.361					

On examination of the within group correlation structure (using the group assignment resulting from the model of complete independence, Model 1), we find that both groups exhibit a high correlation between systolic blood pressure(SBP) and diastolic blood pressure(DBP), 0.629 for group 1 and 0.622 for group 2. Within each group, we tested the categorical variables for pairwise independence. The hypothesis of independence of variables was rejected at the 0.01 level for HX and EKG in group 1. As the correlation existed in one group only, and combination of these two categorical variables would require an additional ten parameters to be estimated, we decided not to combine the variables HX and EKG.

The correlation between the two types of blood pressure is incorporated into a new model, in which variables SBP and DBP are grouped together in a partition. We will refer to this model as Model 2. This model is a mixture of two component distributions, each of which is a product of four discrete distributions, six univariate normal distributions, and one bivariate normal distribution (for the blood pressures). The local independence assumption has been weakened only by adding a covariance parameter between the blood pressures in each of the two

clusters. The iteration for fitting Model 2 may be begun either from the Model 1 parameter estimates or from the cluster assignments based on Model 1 or the clinical classification. Using the Model 1 estimates as starting values, the loglikelihood converged to -11268.723. The algorithm was applied 18 more times, where the initial classifications were generated by randomly allocating the patients into a group. Two other solutions of the likelihood equation were also found. From eighteen starting values, ten converged to a solution where the loglikelihood was -11268.723, six converged to a solution where the loglikelihood was -11275.551, and two converged to a solution where the loglikelihood was -11358.818. The solution corresponding to the largest of the local maxima was the same solution that was found using initial parameters based on the clinical classification or the Model 1 cluster assignments.

When the observations are assigned to their class of greater posterior probability under Model 2, a slightly different classification results from that given by Model 1, as indicated in Table 3.4.

Table 3.4

Agreements and differences between the clinical and Model 2 classifications.

		Model Classification	
	Class	**Group 1**	**Group 2**
Clinical Classification	**Stage 3**	252	21
	Stage 4	21	181

Table 3.4 indicates that there are now 42 differences between the clinical and the model classifications. When we examine the observations that Model 2 classifies in a different group to the clinical classification, and compare these observations with those that are classified differently in Model 1, we find that there is a change in the classification of one observation only. The posterior probabilities of being assigned to Group 1 are listed in Table 3.7 for those observations that change classification under Models 1 and 2, and the yet-to-be-defined Models 3 and 4. Using the same criteria as previously for definite assignment to a group, the same 15 observations that were definitely assigned to a different group than the one equivalent to the clinical classification under Model 1, are also definitely assigned to a group that differs from the clinical classification under Model 2.

The cluster assignment from Model 2 was used, and the within group correlation structure was re-examined. There is a small correlation between Wt, and SBP and DBP, 0.169 and 0.187 for group 1, and 0.166 and 0.262 for group 2. A small correlation also shows up between Wt and HG, 0.193 for group 1, and 0.297 for

group 2. We will fit two further models and introduce both of these correlations separately.

The two group mixture model is fitted with the variables Wt, SBP and DBP grouped together in a partition (Model 3). When the observations are assigned to their class of greater posterior probability under Model 3, a slightly different classification results from that given by Model 2, as indicated in Table 3.5.

Table 3.5

Agreements and differences between the clinical and Model 3 classifications.

		Model Classification	
	Class	**Group 1**	**Group 2**
Clinical Classification	**Stage 3**	252	21
	Stage 4	18	184

As Table 3.5 indicates, there are 39 differences between the statistical diagnosis and the clinical classification. When we examine the observations that the model classifies in a different group to the clinical classification, and compare these observations with those classified differently in Model 2, we find that there is a change in the classifications of five observations only. (Refer to Table 3.7 and the discussion following this table).

The two group mixture model is then fitted with the variables Wt and HG grouped in one partition, and SBP and DBP grouped together in another partition. This model will be referred to as Model 4. The classifications of the observations to their class of greater posterior probability under Model 4 are given in Table 3.6.

Table 3.6

Agreements and differences between the clinical and Model 4 classifications.

		Model Classification	
	Class	**Group 1**	**Group 2**
Clinical Classification	**Stage 3**	252	21
	Stage 4	19	183

Table 3.6 indicates that under Model 4 there are 40 differences between the clinical and the model classification. An examination of the observations that the model classifies in a different group to the clinical classification was made, and these observations were compared with those classified differently in Model 2. It can be seen from Table 3.7 that there has been a change in the classification of two observations only.

Table 3.7

Posterior probabilities for membership in Group 1 for the observations that change classification under Model 1 to Model 4.

		Observation				
		32	58	220	294	482
Model 1	$\hat{z}_{i1}$	0.643	0.582	0.569	0.491	0.489
Model 2	$\hat{z}_{i1}$	0.583	0.625	0.572	0.492	0.519
Model 3	$\hat{z}_{i1}$	0.311	0.432	0.499	0.505	0.446
Model 4	$\hat{z}_{i1}$	0.398	0.517	0.518	0.447	0.421
Clinical Classification		4	3	4	3	4

An investigation of the overall differences in assignment of the observations using Model 1 to Model 4, showed that there was a change in the group assignment of five observations only under the models fitted. It can be seen from Table 3.7 that these observations are not really decisively assigned to any group. Observation 32 is assigned more convincingly to Group 2 under Model 3 and Model 4, whereas under Model 1 it is more convincingly assigned to Group 1. Observation 58 is more convincingly assigned to Group 1 under Model 2 whereas the assignment is not decisive under Model 1, Model 3 and Model 4. The model classifications have been little affected by the choice of model. In practice, this means that it does not really matter which of the four models is selected for this data set, as we will get similar results from each model because of the stability of the group structure. From a clustering viewpoint, the groups formed are very stable under the changes to the model.

The observations whose clinical classification and model group differ for any of the models fitted were examined. Cause of death and treatment group may give us some insight into this difference. Analogously to Byar and Green (1980), the survival status variable was recoded to 4 levels, alive(0), death from prostatic cancer(1), death due to cardiovascular causes(2), and death from other causes(3). The posterior probabilities for being assigned to Group 1, (the less-seriously ill group), the treatment group and the survival status are displayed in Table 3.8 for those observations where the clinical classification and the model group differ for any of the models fitted. It can be seen that in total, there are 43 observations whose model group and clinical classification disagree for any of the models fitted. The model classifications emerge as being very similar to the clinical classification.

It can be seen from Table 3.8 that there are approximately the same proportions of observations that have a clinical classification of Stage 3 and Stage 4, and that have

been assigned by at least one of the models fitted to a different group than the one corresponding to the clinical classification. There are fifteen observations (marked in Table 3.8 with †), that are definitely assigned by all models to a different group than the one equivalent to the clinical classification. It can also be seen that there are more Stage 3 patients than Stage 4 patients, who are definitely assigned by all models fitted, to a different group than the one corresponding to the clinical classification. It is interesting to see from Table 3.8 that virtually all patients that have posterior probabilities $\hat{z}_{i1}$, of less than 0.7, do not survive until the end of the trial.

Table 3.8

Posterior probabilities, Treatment and Survival Status for those observations where the model classification and the clinical classification differ under some of the models fitted.

Obs.	Model 1 $\hat{z}_{i1}$	Model 2 $\hat{z}_{i1}$	Model 3 $\hat{z}_{i1}$	Model 4 $\hat{z}_{i1}$	Stage	Treat -ment‡	Surv. Status
276†	0.000	0.000	0.000	0.000	3	E	1
202†	0.000	0.000	0.000	0.000	3	E	1
181†	0.000	0.000	0.000	0.000	3	E	2
132†	0.000	0.000	0.000	0.000	3	E	1
371†	0.001	0.001	0.001	0.001	3	E	0
345†	0.003	0.003	0.003	0.003	3	E	1
190†	0.009	0.010	0.010	0.008	3	E	2
348†	0.017	0.019	0.018	0.017	3	P	0
273†	0.023	0.025	0.019	0.025	3	P	1
424†	0.028	0.029	0.027	0.023	3	P	3
354†	0.026	0.029	0.026	0.034	3	P	3
272	0.064	0.071	0.072	0.077	3	P	2
274	0.072	0.079	0.089	0.072	3	E	1
502	0.092	0.098	0.105	0.075	3	E	3
356	0.098	0.107	0.105	0.106	3	P	1
99	0.218	0.201	0.186	0.178	3	E	2
246	0.202	0.235	0.256	0.243	3	P	2
124	0.227	0.249	0.250	0.264	3	P	1
353	0.367	0.385	0.365	0.320	3	P	1
325	0.376	0.395	0.413	0.381	3	P	3
482	0.490	0.519	0.446	0.421	4	P	1
294	0.491	0.492	0.505	0.447	3	E	1
32	0.643	0.583	0.311	0.398	4	E	3
192	0.543	0.536	0.523	0.539	4	P	1
58	0.582	0.625	0.432	0.517	3	P	1
220	0.569	0.572	0.499	0.518	4	P	2
37	0.647	0.603	0.615	0.596	4	P	3
197	0.631	0.644	0.677	0.661	4	E	3
221	0.713	0.671	0.654	0.668	4	E	2
392	0.716	0.733	0.735	0.712	4	P	0
458	0.748	0.767	0.768	0.750	4	E	0
65	0.778	0.787	0.771	0.750	4	P	1
427	0.807	0.817	0.801	0.770	4	P	0
457	0.825	0.822	0.814	0.784	4	P	0
462	0.804	0.798	0.823	0.850	4	P	2
117	0.847	0.840	0.836	0.791	4	E	0
114	0.846	0.859	0.862	0.863	4	E	3
116	0.867	0.878	0.871	0.863	4	P	2
196	0.898	0.914	0.916	0.909	4	E	0
396†	0.972	0.972	0.973	0.973	4	P	3
218†	0.984	0.983	0.982	0.981	4	P	3
500†	0.991	0.991	0.981	0.987	4	E	0
393†	0.996	0.996	0.997	0.996	4	P	1

† *Observation definitely assigned by all models to a different group than the one equivalent to the clinical classification.*

‡ E *denotes Estrogen treatment,* P *denotes Placebo treatment.*

How much better do the models with intra-cluster covariances fit? Compared with Model 1, Model 2 has 2 extra parameters - one covariance between blood pressures for each of two clusters - and from Table 3.9 we can see that twice the difference in loglikelihoods is 235.1, clearly a definite improvement. Model 3 adds 4 extra parameters to Model 2 for a $-2\log\lambda$ of 28.0. Model 4 adds covariances between Wt and HG to Model 2 gaining a $-2\log\lambda$ of 29.3 at a cost of 2 parameters. Both Model 3 and Model 4 offer better fitting models than the fully locally independent model for a modest number of extra parameters. The addition of too many covariance parameters is not recommended for fear of upsetting the stability of the model classifications. Model 3 is preferred on physical grounds because we would expect correlations between patient weight and the two blood pressures.

Table 3.9

Loglikelihoods for the four two-group models.

Model	loglikelihood
1	-11386.265
2	-11268.723
3	-11254.743
4	-11254.091

3.2.2 Choosing the number of groups.

As indicated in section 1.5, this thesis uses the likelihood ratio test merely as a guide to the possible number of underlying groups in the mixture. Another guide can be found in the estimates of the posterior probabilities of group membership. Clearly a solution where observations are clearly assigned to a particular component will be of more practical use than one in which many observations have appreciable probability of membership in each of several classes. However it must be remembered that real populations do overlap, and such solutions are not necessarily meaningless.

A comparison between the model fitted under the assumption of a single population ($K = 1$) and the analogous model under the assumption of $K = 2$ groups, gave $-2\log\lambda = 823.2$. We can be confident that there is not a single population. A comparison of the model fitted with two groups and the equivalent model with three groups gave $-2\log\lambda = 188.3$. Twice the difference in the loglikelihood ratio for the models fitted with three and four groups was 175.8. For convenience these values are displayed in Table 3.10. These are all significant χ^2 values. As more groups were included in the model, there seemed to be an increasing tendency to converge to a suboptimal local maxima. This was not unexpected, since each

additional group requires an additional set of 28 parameters to be estimated. We are confident that the best endpoint was reached for the 2 group solution, fairly sure for the 3 group solution, but are not confident for the 4 group solution. We are unsure of whether likelihood singularities are possible with these models, but no instances were encountered where the algorithm failed to converge in the sense of our criterion. For reasons of time it was not practical to investigate 5 group models as the number of possible model varients coupled with increased sensitivity to starting values would make this a lengthy task.

Table 3.10

Likelihood ratio test statistic for K versus $K+1$ clusters.

K	$-2\log\lambda$
1	823.2
2	188.3
3	175.8

Table 3.11

Posterior probabilities for 2-4 Groups.

	No. of Groups		
$\hat{z}_{ij}$	**2**	**3**	**4**
.20 – .80	33	97	140
.80 – .95	44	100	134
.95 – .99	46	63	84
.99 – 1.0	352	215	117

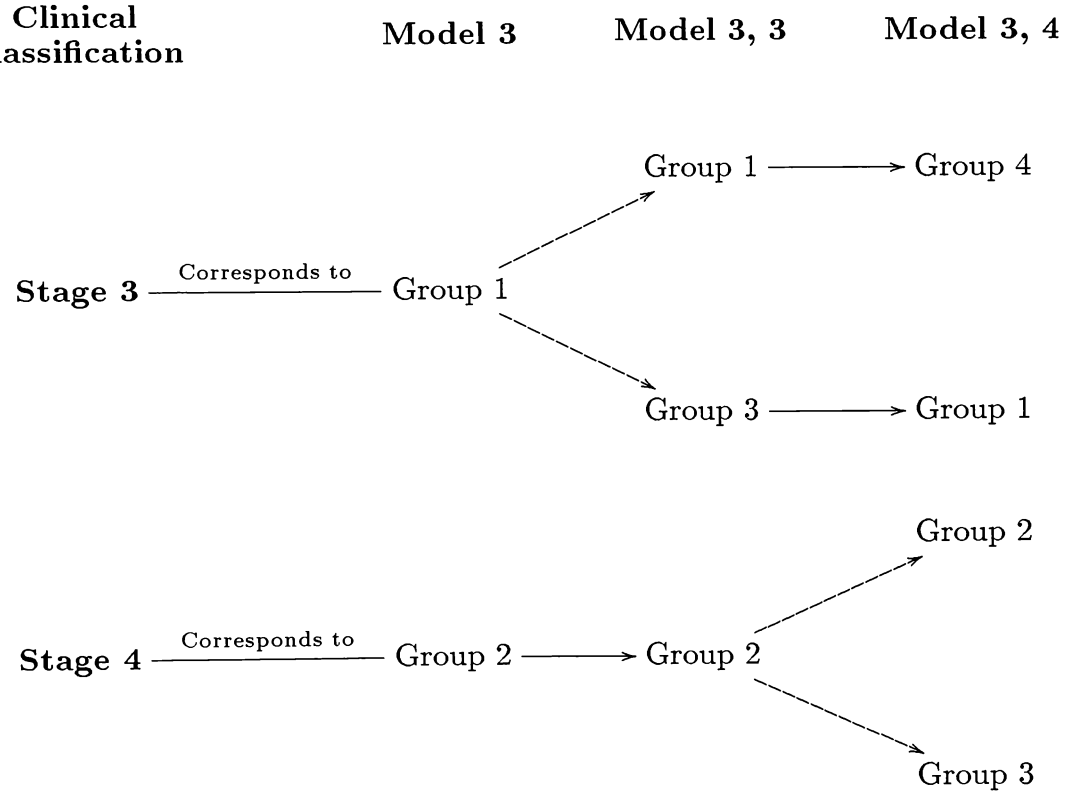
An examination of the posterior probabilities for the groups fitted, found (Table 3.11) that as the number of groups fitted to the data increased, there was a decrease in the number of observations that are definitely assigned to a group ($\hat{z}_{ij} \geq 0.95$). The two cluster model gives more observations that are definitely assigned to a group. In this analysis, the two clusters have an interpretation that agrees with the clinical classification of Stage 3 and Stage 4.

The differences between the assignments of the observations for the 2 cluster and the 3 cluster models was investigated. When a three group model was fitted, most (96%) of the Group 2 (the more seriously ill), patients were assigned to a single cluster, whilst the bulk (98.5%) of the Group 1 (the less seriously ill), patients were divided between the 2 other clusters. The differences between the assignments of the observations for the three group and the four group models were also investigated. It was found for the four group model that the group assignment

for the two groups corresponding to the less seriously ill patients was virtually identical for the three and four group models, however the group corresponding to the more seriously ill in the three group model (AND Model 3) had been divided between two other clusters. This type of division indicates the stability of the grouping structure under the models fitted. The relationships between the groups for the models fitted are illustrated in Figure 3.1.

Figure 3.1

Relationships between the groups under the models fitted



3.2.3 Clusters and outcomes

Additional insight into the composition of the groups may be gained from examining the cause of death. For convenience, the recoded survival status variable is listed again;- alive (0), death from prostatic cancer(1), death due to cardiovascular causes(2), and death from other causes(3).

Table 3.12

Survival Status for Model 3 Classifications.

Group	Survival Status			
	0	1	2	3
1	96	24	92	58
2	41	97	46	21

We can see (Table 3.12) that patients in Group 1 (corresponding to the clinical classification of Stage 3) have a higher probability of being alive or dying from cardiovascular causes, whereas patients in Group 2 (clinical classification of Stage 4) have more chance of dying from prostatic cancer.

Table 3.13

Survival Status for Model 3, 3.

Group	Survival Status			
	0	1	2	3
1	56	18	31	21
2	38	91	44	18
3	43	12	63	40

In Table 3.13, we consider the survival status for Model 3, 3, the three group model with the same partitioning as for Model 3. Group 2 for this model corresponds roughly to the clinical classification of Stage 4 (Group 2 in the 2 cluster solution). The patients in the Group 1 have a higher probability of being alive at the end of the trial whereas the Group 3 patients have a higher probability of death from cardiovascular causes, and similar moderate probabilities of death from other causes, and alive at the end of the trial.

Table 3.14

Survival Status for Model 3, 4.

Group	Survival Status			
	0	1	2	3
1	45	10	62	41
2	29	40	26	13
3	12	51	20	9
4	51	20	30	16

In Table 3.14, the survival status is considered for Model 3, 4, the four group model with the same partitioning as in Model 3. The patients in Group 2 have a higher probability of death from prostatic cancer and approximately the same probability of being alive at the end of the trial, and death from cardiovascular causes. The patients in Group 3 also have a higher probability of death from prostatic cancer. Group 3 consists of the more seriously ill Stage 4 patients, while Group 2 consists of the less seriously ill Stage 4 patients. When the results of Table 3.14 are compared with those in Table 3.13, it can be seen that under the four group model, Group 1 corresponds to Group 3 of the three group model, Group 2 and Group 3 (Table 3.14) correspond to Group 2, the more seriously ill patients in Table 3.13, and Group 4 (Table 3.14) corresponds to Group 1 in Table 3.13.

Further insight into the composition of the groups may be obtained from an examination of the survival status for the treatments the patients received. At the time of diagnosis of the cancer, patients had been randomly assigned to one of four treatments:– placebo, $0.2mg$ diethylstilbestrol, (henceforth denoted by DES), $1.0mg$ DES, and $5.0mg$ DES. Byar and Green (1980) noted that previous analyses found that $0.2mg$ DES had a similar effect on the cancer as the placebo had, and that the $1.0mg$ and $5.0mg$ DES treatments also had similar effects on the cancer. Analogous to Byar and Green (1980), the treatments, placebo and $0.2mg$ DES, will be combined into one treatment group, which shall be designated as 'placebo treatment', and the treatments, $1.0mg$ and $5.0mg$ DES will be combined into another treatment group which shall be denoted as 'estrogen treatment'. The group structure will be investigated by looking at the survival status of the patients for each of the two treatment groups, placebo and estrogen.

Figure 3.2 shows that for the Group 1 patients, the less seriously ill, estrogen treatment increased the percentage of patients being alive at the end of the trial, and decreased the percentage of patients dying from both prostatic cancer and other causes. However, estrogen treatment increased the percentage of patients dying from cardiovascular causes. Estrogen treatment would probably be the preferred treatment for patients in this group, provided account was taken of the increased risk of death from cardiovascular causes. With the Group 2 patients, both estrogen treatment and placebo treatment resulted in very similar proportions of patients dying from cardiovascular causes, and from other causes. In comparison to the placebo treatment, estrogen treatment decreased the proportions of patients dying from prostatic cancer and increased the proportions of patients who survived to the end of the trial. Estrogen treatment would also be the preferred treatment for patients in this group.

The treatment effects will now be investigated for Model 3, 3. It can be seen

from Figure 3.3, that similar proportions of the Group 1 patients died from both cardiovascular causes, and other causes for the placebo and estrogen treatments. Estrogen treatment resulted in a large increase in the percentages of patients being alive at the end of the trial, and a large decrease in the percentages of patients dying from prostatic cancer. Estrogen treatment is the preferred treatment for this group. With the Group 2 patients, estrogen treatment again has a beneficial effect by giving a slight increase in the percentage of patients alive at the end of the trial, and a slight decrease in the percentage of patients dying from prostatic cancer. Both estrogen and placebo treatments have similar percentages of patients dying from cardiovascular causes and from other causes. Estrogen treatment has a small advantage over placebo treatment for patients in this group.

Both estrogen and placebo treatment have a similar effect on the proportions of patients who were still 'alive at the end of the trial' for the Group 3 patients. In comparison to those receiving placebo treatment in this group, estrogen treatment decreased the percentage of patients dying from prostatic cancer, and from other causes, but greatly increased the percentages of patients dying from cardiovascular causes.

An investigation of the treatment effects on the group structure for Model 3, 4 was also made. (see Figure 3.4). The treatment effects for Group 1 and Group 4 are very similar to those for the corresponding groups in the three group model. It has previously been shown in Figure 3.1 that in the four group model, there are basically two clusters for the Stage 3 patients and two clusters for the Stage 4 patients. It can be seen that in Group 2, estrogen and placebo treatment have a virtually identical effect on the percentage of patients being alive at the end of the trial and dying from other causes. Estrogen treatment for the patients in this group, results in a decrease in the proportions of patients dying from prostatic cancer, and an increase in the proportions dying from cardiovascular causes. Estrogen treatment is probably the preferred treatment for patients in this group. It can also be seen from Figure 3.4, that estrogen treatment would be the preferred treatment for the Group 3 patients, as it increases the percentage of patients being alive at the end of the trial, and decreases the percentage of patients dying from prostatic cancer, cardiovascular and other causes.

The group structure, treatment effects and survival outcome were looked at with respect to age for Model 3, 4 . It could be seen that younger patients (Groups 3 and 4), fared better with estrogen treatment, whereas with the older patients, estrogen treatment resulted in an increase in 'deaths by cardiovascular causes'. This is similar to the conclusions reached by Byar and Green (1980).

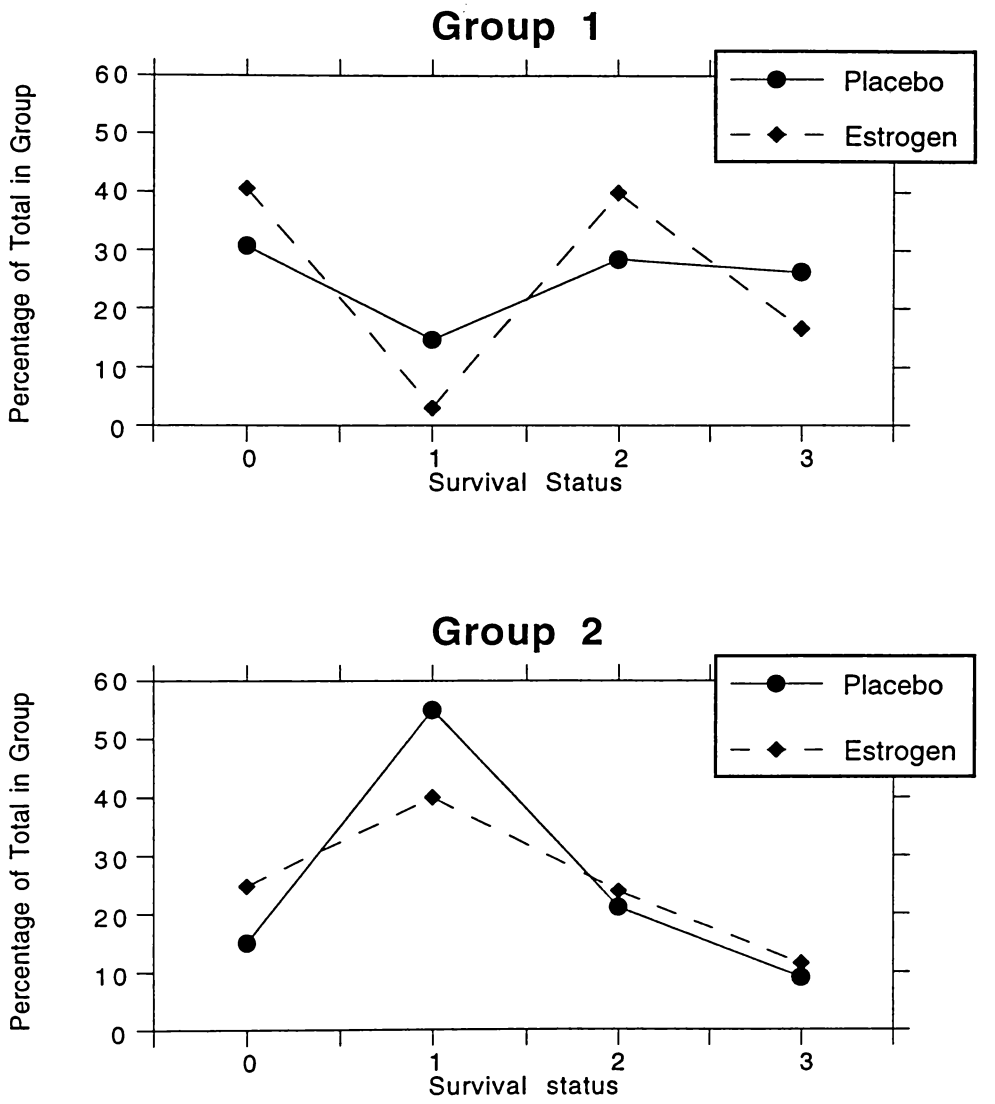


Figure 3.2

Percentage of total in group versus survival status by treatments, Placebo and Estrogen, for Model 3.

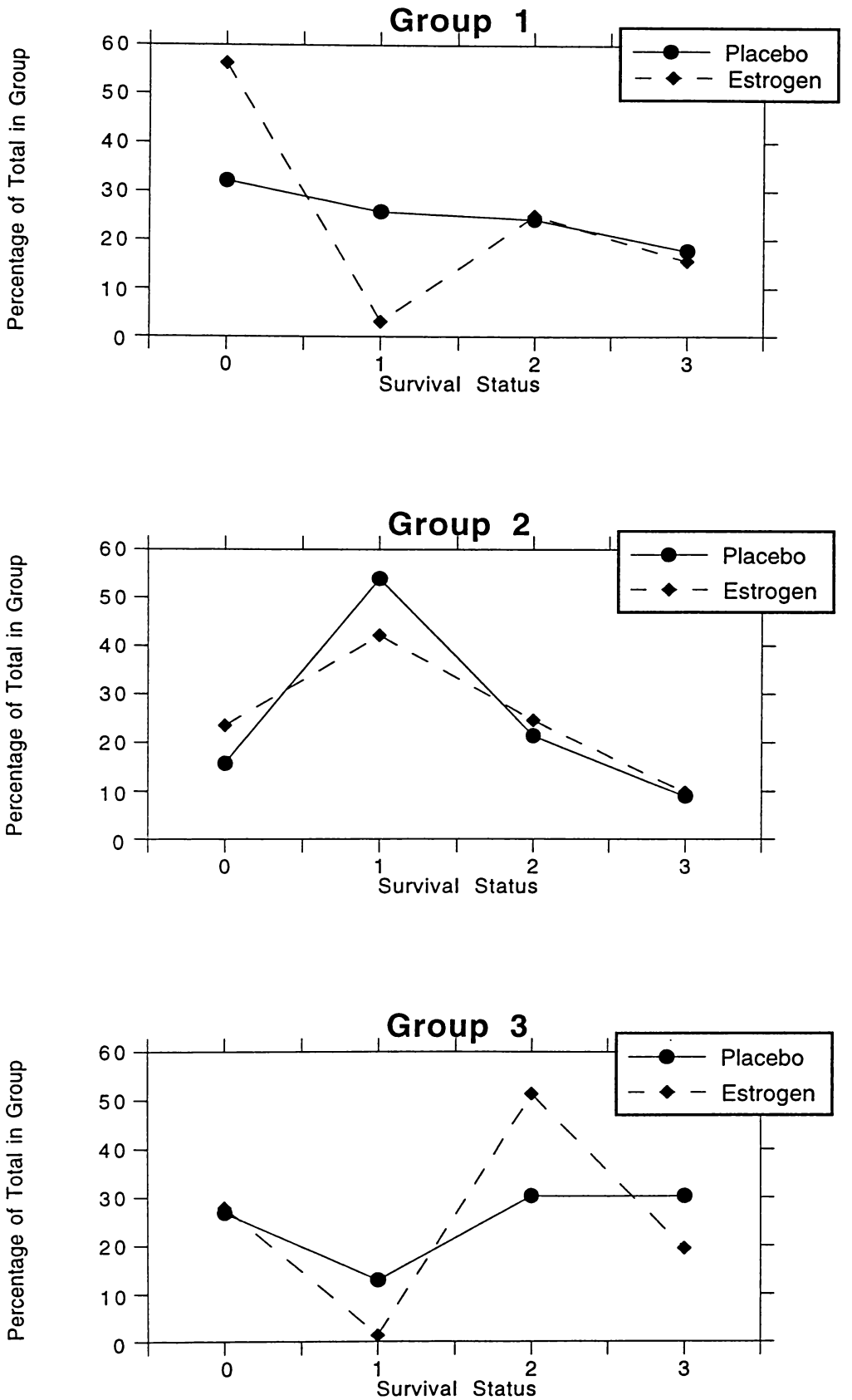
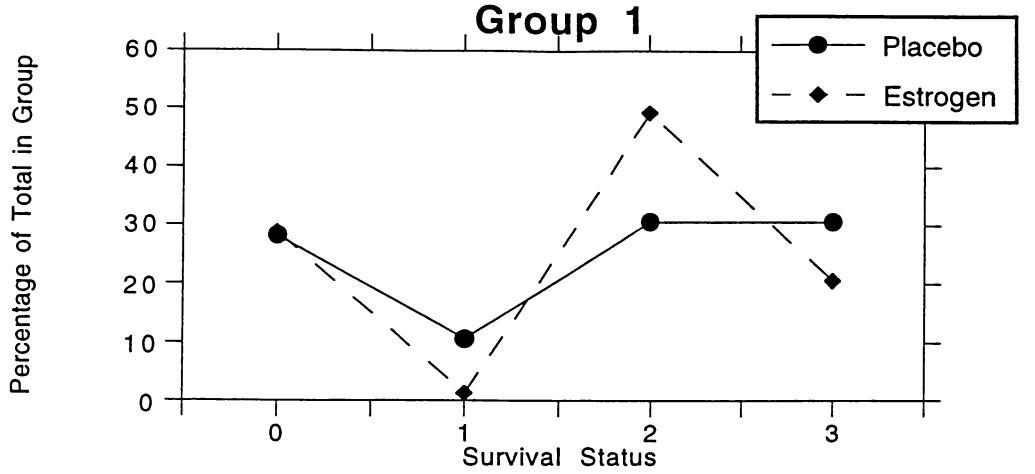
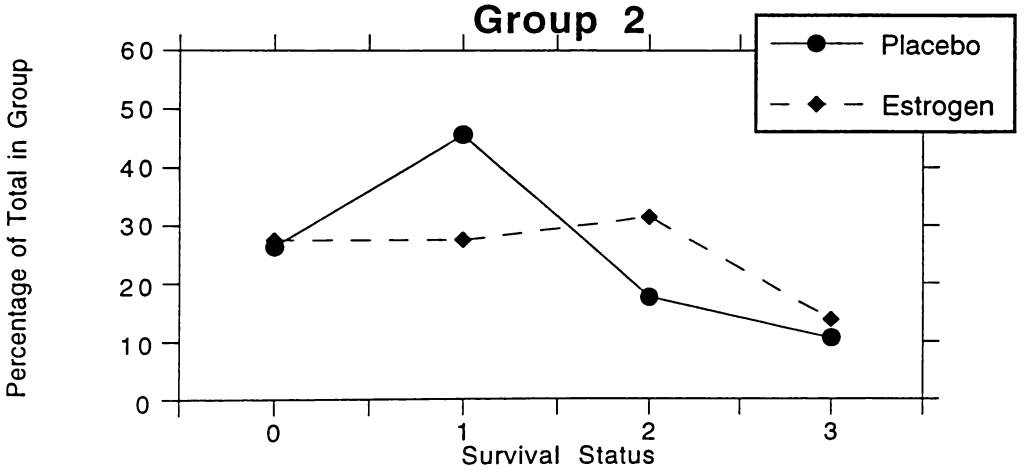
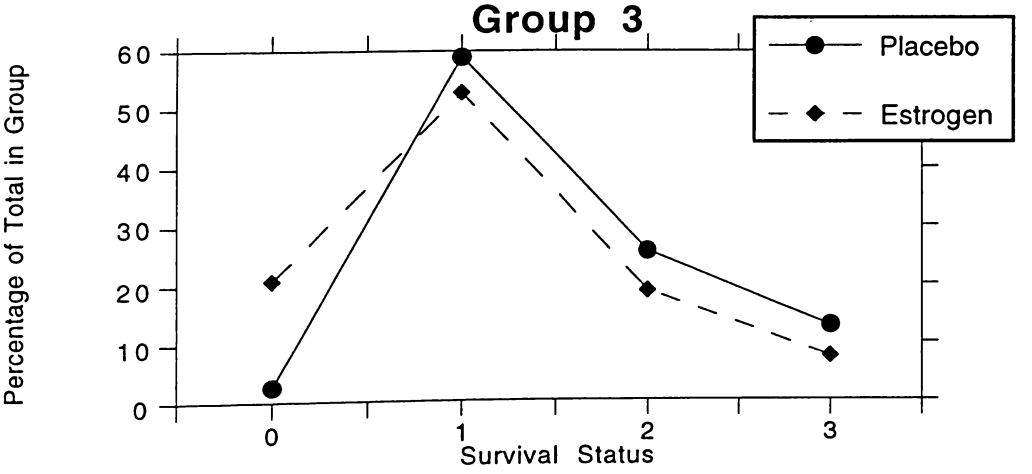


Figure 3.3

Percentage of total in group versus survival status by treatments, Placebo and Estrogen, for Model 3, 3.


Group 1
Placebo
Estrogen
Percentage of Total in Group
60
50
40
30
20
10
0
0
1
2
3
Survival Status


Group 2
Placebo
Estrogen
Percentage of Total in Group
60
50
40
30
20
10
0
0
1
2
3
Survival Status


Group 3
Placebo
Estrogen
Percentage of Total in Group
60
50
40
30
20
10
0
0
1
2
3
Survival Status

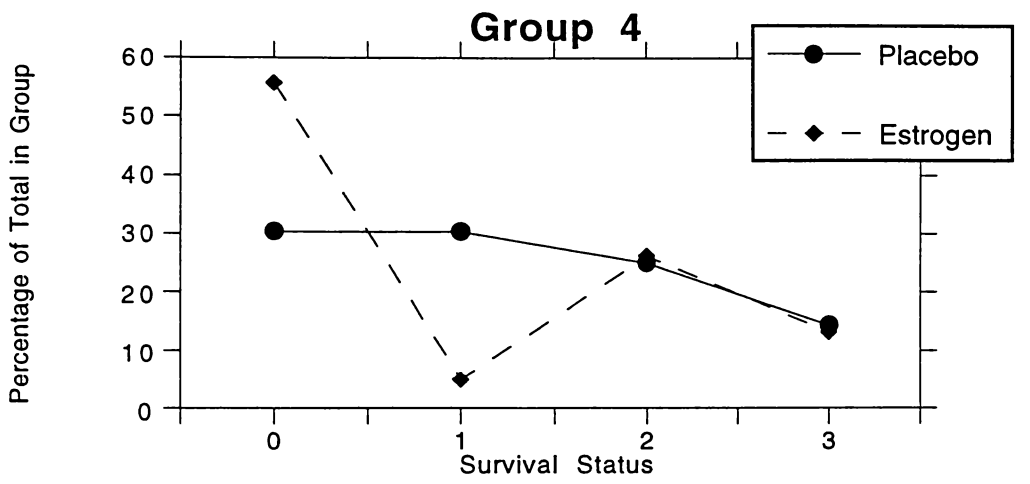


Figure 3.4

Percentage of total in group versus survival status by treatments, Placebo and Estrogen, for Model 3, 4.

The variable 'months of follow-up', provides another way to gain insight into the composition of the group structure as it can be regarded as a surrogate for survival time. Because each patient in the study was followed up for at least four years unless death occurred, this variable was categorized to 'survival time greater than 48 months', and 'survival time less than or equal to 48 months'. The survival time was investigated for the two group model, specifically Model 3. Table 3.15 tabulates this with group assignment.

Table 3.15

Survival Time for Model 3 clusters.

	Survival Time	
Group	$\leq$ 48 months	$>$ 48 months
1	144	126
2	150	55

With the Group 1 patients, 46.7% survive for greater than 48 months, whereas only 26.8% of the Group 2 patients survive for greater than 48 months.

It is intriguing to have a closer look at the patients whose classification by Model 3 is in conflict with the clinical classification, in Table 3.16 we tabulate survival time against the model classification for these patients alone. In Table 3.17 survival status information is summarised for the same patients.

Table 3.16

Survival time for the observations classified by the model to a different group than the clinical classification.

	Survival Time	
Group	$\leq$ 48 months	$>$ 48 months
1	9	9
2	18	3

Table 3.17

Survival status for the observations classified by Model 3 to a different group than the clinical classification.

	Survival Status			
Group	0	1	2	3
1	7	3	3	5
2	2	10	5	4

Tables 3.16 and 3.17 suggest a more favourable outcome for Stage 4 patients classified by the model into Group 1 than for Stage 3 patients classified into Group 2. In short, the model classification gives a better indication of prognosis than the clinical classification, the patients in Group 2 being likely to succumb to prostatic cancer, and the patients in Group 1 more likely to survive, or die from other causes.

The investigation has shown that the clusters found using the approach of multivariate finite mixture models have real meaning. The groups that were found when using multivariate finite mixture models with two groups, are basically equivalent to the groups into which the patients have been classified by clinical criteria. When a three group model was fitted to the data, the Stage 4 patients basically stayed in one group, whilst the majority of the Stage 3 patients were divided into two groups, one group containing patients who were more likely to die from cardiovascular causes, and the other group containing those who were more likely to be alive at the end of the trial. When a four group model was fitted to the data, the two groups containing the bulk of the Stage 3 patients were virtually identical to those found for these patients in the three group model, whilst the Stage 4 patients had been divided in two groups, one of which was more seriously ill than the other. We can conclude that the approach of multivariate finite mixture models has successfully clustered a large complex data set.

3.3 Fishers Iris data

We will now consider the well known iris data first discussed by Fisher (1936). This data set consists of four variables, sepal length, sepal width, petal length and petal width, measured on 50 plants from each of three species of iris, *Iris Setosa*, *Iris Versicolor* and *Iris Virginica*. From previous analyses, see for example, Krzanowski (1988), Mardia, Kent, and Bibby (1979), and Morrison (1976), we know that the three species of iris are widely different, but that *I. Versicolor* and *I. Virginica* are more similar to each other than either is to the *I. Setosa*. McLachlan (1992) in chapter 6, used this data to 'test' for the existence of subspecies in both the *I. Versicolor* and *I. Virginica* species.

This data set has been analysed for the purpose of comparing some of the results obtained from using the approach of multivariate finite mixture models with those of Everitt and Mérette (1990). In the following analyses, we will only consider models with three clusters. The extent of recovery of the grouping structure known to exist in this data set will then be evaluated by computing the Hubert and Arabie (1985) adjusted rand index for each model fitted. The strategy that has previously been used for fitting models will be followed: firstly, the model for complete local independence is fitted; then, using the group assignment from the model for complete independence, the within group correlation structure will

be examined. Variables with strong within cluster associations will be grouped together in a partition cell for the next series of fits. This process may be repeated.

A preliminary examination of the data indicated that all four variables were approximately normally distributed in the mixture population, though the petal length and petal width were bi-modal distributions. A mixture of three components with separate variances was fitted to see the extent to which the model could rediscover the species.

3.3.1 Three component models

We regard the data as a random sample from the distribution

$$f(\mathbf{x};\boldsymbol{\phi}) = \sum_{k=1}^{3} \pi_k f_k(\mathbf{x};\boldsymbol{\theta}_k)$$

where $\sum_{k=1}^{3} \pi_k = 1$, and $\pi_k \geq 0$, $k = 1,\ldots,3$. Under the model with complete local independence for three clusters, the component distributions will be of the form

$$f_k(\mathbf{x}_i;\boldsymbol{\theta}_k) = \prod_{l=1}^{4} f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl}),$$

where $\boldsymbol{\theta}_{kl}$ is the parameter vector for group k, partition cell l, and $k = 1,\ldots,3$, $l = 1,\ldots,4$. We see that $f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl})$ is the $N(\mu_{kl},\sigma^2_{kl})$ density function for each of the four continuous variables.

The notation that has previously been defined in section 3.2.1 will used. However, to distinguish between models for the cancer data and models for the iris data, 'Iris' will be included in the description of all models for the iris data. As we will only be investigating models that have three groups, the number of groups to be fitted to the model will be omitted from the notation. Hence, 'Iris Model i' denotes the model for the iris data that has partitioning i and three groups. We will refer to the model with complete local independence as Iris Model 1.

The model was fitted iteratively using the EM algorithm, with the initial estimates of the group parameters being based on those resulting from the three species. In order to search for local maxima, the EM algorithm was run fifteen more times from initial classifications generated by randomly splitting the observations into three groups of size fifty. Two solutions of the likelihood equation were found for Iris Model 1. From fifteen starting values, thirteen converged to a solution with a loglikelihood of -306.860, the same solution that was found using the parameters based on the species classification. Two starting values converged to a solution with a loglikelihood of -307.178. The solution corresponding to the larger of the two local maxima was taken. As previously, an observation $\mathbf{x}_i$ is assigned to the population to which it has the highest estimated posterior probability of belonging.

Table 3.18

Agreements and differences between the species and the model classifications for Iris Model 1.

Species	Classification Model		
	Group 1	**Group 2**	**Group 3**
I. Setosa	50	0	0
I. Versicolor	0	43	7
I. Virginica	0	2	48

It can be seen from Table 3.18, that under Iris Model 1, the species classification and the model classification are different for 9 plants. Group 1 corresponds to *I. Setosa*, Group 2 is predominantly *I. Versicolor* whilst Group 3 is formed from overlapping clusters of the Versicolor and the Virginica species. All *I. Setosa* plants are assigned to Group 1 with posterior probabilities of $\hat{z}_{i1} = 1.00$, whilst all Versicolor and Virginica plants have posterior probabilities of being assigned to Group 1 with $\hat{z}_{i1} = 0.00$. Iris Model 1 has also detected that the Versicolor and the Virginica species overlap whilst the Setosa species is quite separate to the others.

We shall speak of an observation as being definitely assigned to a group when the posterior probability $\hat{z}_{ik}$ of assignment to Group k is greater than or equal to 0.95. On examination of the posterior probabilities for Iris Model 1, we find that 132 of the 150 observations are definitely assigned to a group. Of the observations that are misclassified, four are definitely assigned to a different group. Observations 71 and 78 both belong to species *I. Versicolor* and are definitely assigned to Group 3 with $\hat{z}_{71,3} = .9596$ and $\hat{z}_{78,3} = .9862$, whilst observations 107 and 120 are both *I. Virginica* and are definitely assigned to Group 2 with $\hat{z}_{107,2} = .9854$ and $\hat{z}_{120,2} = .9740$.

Another comparison between the Iris Model 1 fit and the species classification can be obtained by comparison of the estimated parameters under Iris Model 1 with their counterparts under the species classification. This is done in Table 3.19, and it can be seen that the estimates are quite close. The estimated proportions in each cluster are 0.333, 0.305, and 0.362 leading to expected numbers in each cluster of 50.0, 45.8, and 54.2. This is fairly close to the 50 observations in each species.

Table 3.19

Summary statistics of the four variables according to the species and the Iris Model 1 estimates

		Classification				
		Species			Model	
Variable	**Species**	**Mean**	**STDEV**	**Group**	**Mean**	**STDEV**
Sepal length	*I. Setosa*	5.006	0.353	1	5.006	0.349
	I. Versicolor	5.936	0.516	2	5.835	0.478
	I. Virginica	6.588	0.636	3	6.623	0.570
Sepal width	*I. Setosa*	3.428	0.379	1	3.428	0.375
	I. Versicolor	2.770	0.313	2	2.700	0.295
	I. Virginica	2.974	0.322	3	3.017	0.288
Petal length	*I. Setosa*	1.462	0.174	1	1.462	0.172
	I. Versicolor	4.260	0.470	2	4.222	0.475
	I. Virginica	5.552	0.552	3	5.483	0.572
Petal width	*I. Setosa*	0.246	0.105	1	0.246	0.104
	I. Versicolor	1.326	0.198	2	1.304	0.187
	I. Virginica	2.026	0.275	3	1.990	0.292

We will seek ways of improving the agreement between the species and the model classifications by adding in more covariance parameters. The group assignment from Iris Model 1 was used and the within group correlation structure was examined.

Table 3.20

Within group correlation structure using the group assignment from Iris Model 1. (lower triangle only).

Group 1.

	Sepal length	Sepal width	Petal length
Sepal width	0.743		
Petal length	0.267	0.178	
Petal width	0.278	0.233	0.332

Group 2.

	Sepal length	Sepal width	Petal length
Sepal width	0.419		
Petal length	0.662	0.375	
Petal width	0.375	0.477	0.731

Group 3.

	Sepal length	Sepal width	Petal length
Sepal width	0.373		
Petal length	0.759	0.226	
Petal width	0.178	0.326	0.432

Using a significance level of 0.05, the critical values for a test of $H_0 : \rho = 0$ are $r^* = 0.278$, 0.294 and 0.266 for sample sizes equal to the numbers assigned to Group 1, Group 2 and Group 3 respectively. It can be seen from Table 3.20 that there are apparently significant correlations between each pair of variables in at least one of the groups. All 4 variables were therefore grouped together into one partition cell, allowing for a general correlation pattern.

We will fit a mixture of three component distributions, each of which consists of a multivariate normal distribution with 4 variables. The component distributions will be of the form

$$f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) = f_{k1}(\breve{\mathbf{x}}_{i1}; \boldsymbol{\theta}_{k1}),$$

where $\boldsymbol{\theta}_{k1}$ is the parameter vector for group k, partition cell 1, and $k = 1, \ldots, 3$. We see that $f_{k1}(\breve{\mathbf{x}}_{i1}; \boldsymbol{\theta}_{k1})$ is the $N_4(\boldsymbol{\mu}_{k1}, \Sigma_{k1})$ density. We will refer to this model as Iris Model 2.

The model was fitted using the parameter estimates based on those resulting from the species classification. Using these estimates as starting values, the loglikelihood converged to -180.185. In order to search for local maxima, the EM algorithm was applied another sixteen times from initial classifications generated by

randomly splitting the observations into three groups. Four solutions of the likelihood equation were found. Eight starting values converged to a solution where the loglikelihood was -180.185, the same solution that was found using the parameter estimates based on the statistics from the species classification. Four starting values converged to a solution where the loglikelihood was -186.569, two converged to a solution where the loglikelihood was -189.503 and two converged to a solution where the loglikelihood was -192.655. The solution corresponding to the largest of the local maxima was used.

When the observations are assigned to their group of greatest posterior probability under Iris Model 2, a slightly different classification results from that given by Iris Model 1.

Table 3.21

Agreements and differences between the species and the model classifications for Iris Model 2.

Species	Classification Model		
	Group 1	**Group 2**	**Group 3**
I. Setosa	50	0	0
I. Versicolor	0	45	5
I. Virginica	0	0	50

Table 3.21 indicates that there are now five differences between the species and the model classifications. Group 1 corresponds to the species classification of *I. Setosa*, Group 2 is comprised of 90% of the *I. Versicolor* plants, whilst Group 3 consists of all the *I. Virginica* plants and some of the *I. Versicolor*. The observations that the model classifies differently all belong to *I. Versicolor*, and are classified by Iris Model 2 as Group 3.

The posterior probabilities for Iris Model 2 were examined, and under this model, it was found that 141 observations are now definitely assigned to a group. The posterior probabilities are listed in Table 3.22, for the five *I. Versicolor* plants that are classified under Iris Model 2 into the same group as the *I. Virginica* plants.

Table 3.22

Posterior probabilities for the I. Versicolor plants classified into the same group as the I. Virginica plants under Iris Model 2.

Obs.	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{z}_{i3}$
69	0	0.00275	0.99725
71	0	0.05268	0.94732
73	0	0.04156	0.95844
78	0	0.32860	0.67140
84	0	0.00671	0.99329

It can be seen from Table 3.22 that observation 78 is not decisively assigned to either Group 2 or Group 3, whereas the other observations are definitely assigned to a different group than the species classification.

The improvement in fit gained by adding covariances to Iris Model 1 was investigated. Compared with Iris Model 1, Iris Model 2 requires an additional 18 parameters to be estimated, (6 parameters for each group). Twice the difference in the loglikelihoods is 253.350. There is clearly a definite improvement from using Iris Model 2 in preference to Iris Model 1. This can also be seen from the increase in the number of observations that are definitely assigned to the groups.

However, when using Iris Model 2, we are estimating a total of forty four parameters from quite a small data set (150 observations, with 4 variables measured on each observation). Although the test for equality of variance covariance matrices was rejected, it was decided that due to the large number of parameters being estimated, we would fit the model in which all four variables were grouped in one partition cell, and which assumed equal covariance matrices for the component distributions. This model will be referred to as Iris Model 3. The component distributions $f_{k1}(\tilde{\mathbf{x}}_{i1}; \boldsymbol{\theta}_{k1})$, for this model, will be of the form $N(\boldsymbol{\mu}_k, \Sigma)$ for each of the three groups.

The program written for this thesis, in its current version, does not allow for equal covariance matrices to be imposed for the component distributions. The following results for Iris Model 3 were obtained using the published program KMM of Basford and McLachlan (1988). The model was fitted with the initial estimates of the group parameters being based on those resulting from the species classification. In order to search for local maxima, the EM algorithm was run another ten times from initial classifications generated by randomly splitting the observations into three groups. Two solutions of the likelihood equation were found for Iris Model 1. Six starting values converged to a solution where the loglikelihood was

-256.354, the solution that was found using the parameters based on the species classification. Five starting values converged to a solution where the loglikelihood was -263.474. The solution corresponding to the larger of the two local maxima was used.

Table 3.23

Agreements and differences between the species and the model classifications for Iris Model 3.

Species	Classification Model		
	Group 1	**Group 2**	**Group 3**
I. Setosa	50	0	0
I. Versicolor	0	48	2
I. Virginica	0	1	49

It can be seen from Table 3.23 that there are three differences between the model and the species classification. An examination of the posterior probabilities for Iris Model 3 showed that 137 observations are now definitely assigned to a group. The three observations, plants 71, 84 and 134, that are classified to a different group than the one corresponding to the species classification have posterior probabilities respectively of $\hat{z}_{71,2} = 0.133$, $\hat{z}_{84,2} = 0.127$, and $\hat{z}_{134,2} = 0.745$. (Note that $\hat{z}_{i,1} = 0.00$ for these observations). It can be seen that these observations are not definitely assigned to either Group 2 or Group 3.

Assumption of a common covariance structure for the iris data, has improved the agreement between the model and the species classifications. There are now fewer observations definitely assigned to a group than there were under Iris Model 2.

As expected the cluster corresponding to *I. Setosa* is stable under the three models fitted. Both *I. Versicolor* and *I. Virginica* species always had posterior probabilities of 0.00 of being assigned to the same group as *I. Setosa* plants, whilst *I. Setosa* plants were all definitely assigned to the Group 1 with posterior probabilities of 1.00. This occurred for all three models fitted.

In the next subsection, additional models are fitted to this data. We do not recommend this procedure to be followed as a general rule. This procedure has been followed for comparison purposes only.

3.3.2 Categorisation of two continuous variables

Everitt (1988) assumed that the categorical variables in a data set arise from applying thresholds to underlying unobservable continuous variables and that the

continuous variables, both latent and manifest, have a multivariate normal mixture density. Further details can be found in Section 1.8. Everitt and Mérette (1990) chose the iris data to illustrate this approach by firstly categorising the two continuous variables, sepal length and sepal width. We will use the same categorisation. Specifically, sepal length was categorised with three possible outcomes: $x \leq 5.4$, $5.4 < x \leq 6.2$ and $x > 6.2$; and sepal width had two outcomes: $x \leq \mu$ (3.428) and $x > \mu$ (3.428).

We will thus fit a mixture of three component distributions, each of which is a product of two discrete distributions and one bivariate normal distribution (for the petal length and petal width). The component distributions are of the form

$$f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) = \prod_{l=1}^{3} f_{kl}(\tilde{\mathbf{x}}_{il}; \boldsymbol{\theta}_{kl}),$$

where $f_{kl}(\tilde{\mathbf{x}}_{il}; \boldsymbol{\theta}_{kl})$ is $N(\boldsymbol{\mu}_{kl}, \Sigma_{kl})$ for the two continuous variables, and $D(\lambda_{kl1}, \ldots, \lambda_{klm_l})$ for each of the two categorical variables[2]. We will refer to this model as Iris Model 4.

The model was fitted using the EM algorithm from initial estimates of the parameters based on those resulting from the species classification, and also from initial classifications generated by randomly splitting the observations into three groups. Three solutions of the likelihood equation were found. From ten starting values, eight converged to a solution where the loglikelihood was -292.709, the solution that was found using the parameters based on the species classification. One starting value converged to a solution where the loglikelihood was -293.243, and one converged to a solution where the loglikelihood was -307.69386. The solution corresponding to the largest of the three maxima was used. When the observations were assigned to their group of greatest posterior probability under Iris Model 4, the composition of the groups has changed greatly from those given in previous models.

Table 3.24

Agreements and differences between the species and the model classifications for Iris Model 4.

Species	Classification Model		
	Group 1	**Group 2**	**Group 3**
I. Setosa	50	0	0
I. Versicolor	0	37	13
I. Virginica	0	10	40

[2] refer to notation in Chapter 2.

Table 3.24 indicates that there are now 23 differences between the species and the model classification. The species *I. Versicolor* and *I. Virginica* are more difficult to separate using Iris Model 4, with Group 2 and Group 3 both being formed from overlapping clusters of *I. Virginica* and *I. Versicolor*. On examination of the posterior probabilities, we find that 126 observations are definitely assigned to a group under Iris Model 4. An examination of the posterior probabilities of the observations whose species and model classification differ, showed that only seven of these observations that are assigned to a different group to the species are assigned definitely to that group.

It can be seen from Table 3.20 that there is a correlation between the continuous variables, sepal length and sepal width. Hence, it is expected that there may be some association between the two variables that result from categorising sepal length and sepal width. A test of the hypothesis of independence of the categories of these two variables was made for the three species. This test was rejected at the 0.05 level for *I Versicolor* with a $\chi^2_{(2)}$ value of 10.601, and rejected at the 0.06 level for *I Virginica* with a $\chi^2_{(2)}$ value of 5.447. It was not rejected for *I Setosa* ($\chi^2_{(1)} = 1.058$). The two categorical variables were therefore combined into one categorical variable having 6 levels.

A mixture of three component distributions was fitted, where each component consisted of a product of a discrete distribution containing the categorical variable with six levels, and one bivariate normal distribution. We will refer to this model as Iris Model 5. The model was fitted using the EM algorithm using parameter estimates based on those resulting from the species classification, and also from initial classifications generated by randomly splitting the observations into three groups. Two solutions of the likelihood equation were found. From ten starting values, eight converged to a solution where the loglikelihood was -289.023, the same solution that was found using the parameter estimates based on the statistics from the species classification. Two starting values converged to a solution where the loglikelihood -299.136. The solution corresponding to the larger of the two local maxima was used. When the observations were assigned to their group of greatest posterior probability under Iris Model 5, a different classification results from that of Iris Model 4, as is shown in Table 3.25.

Table 3.25

Agreements and differences between the species and the model classifications for Iris Model 5.

Species	Classification Model		
	Group 1	**Group 2**	**Group 3**
I. Setosa	50	0	0
I. Versicolor	0	50	0
I. Virginica	0	12	38

There are now 12 differences between the species and the model classification. The composition of the grouping has changed with 12 of the *I. Virginica* observations being grouped together with the *I. Versicolor* observations. Group 3 now consists of 38 of the *I. Virginica* observations. An examination of the posterior probabilities showed that 131 observations are now definitely assigned by Iris Model 5 to a cluster. Five of the observations that are classified by the model to a different group than the species, are definitely assigned to a different group than the species.

When the observations that Iris Model 5 classifies in a different group to the species are compared with those observations that are classified differently under Iris Model 4, it was found that the same ten *I. Virginica* observations are classified as Group 2 under both models.

It can be seen from a comparison of Table 3.24 and Table 3.25, that the combination of the two categorical variable into one variable having six levels has improved the agreements between the model and the species classifications. There is also an increase in the numbers of observations that are definitely assigned to a group under Iris Model 5. However, Iris Model 5 requires 35 parameters to be estimated from the data, whilst Iris Model 4 requires the estimation of 32 parameters.

3.3.3 Comparison using the Rand index

Everitt and Mérette (1990) calculated an adjusted Rand index for the model they proposed. To compare the results using the approach of multivariate mixture models with those of Everitt and Mérette, the Hubert and Arabie (1985) adjusted Rand index was calculated for all models.

The adjusted Rand index is defined as follows.

$$\text{Adjusted Rand index} = \frac{a + d - n_c}{a + b + c + d - n_c}$$

where a gives the number of pairs of entities that are correctly grouped in the same cluster by true solution and the algorithm solution, d gives the number of pairs of

entities that are correctly grouped in different clusters; b indicates the numbers of pairs of entities that the algorithm has incorrectly placed in the same cluster when the points actually came from different clusters; and c gives the count for the pairs of entities that the algorithm has incorrectly placed in different clusters when they actually came from the same cluster. The adjustment term n_c is defined as

$$\frac{n(n^2+1) - (n+1)\sum n_{i.}^2 - (n+1)\sum n_{.j}^2 + \sum\sum n_{i.}^2 n_{.j}^2/n}{2(n-1)}$$

where n_{ij} is the number of points in cluster i as produced by the algorithm that are also in cluster j of the true solution. Then, $n_{i.}$, $n_{.j}$ and n give the marginals and grand total for such a classification table. The adjustment term has been designed to correct the Rand index for the presence of chance agreement. Milligan and Cooper (1986) concluded that the Hubert and Arabie adjusted Rand index was the best choice for evaluating the extent of recovery of the true structure in a clustering solution.

Table 3.26

Adjusted rand index for the models investigated.

Model	Adjusted Rand Index	Number of Observations Definitely Assigned	Misclassified
Iris Model 1	0.868	132	9
Iris Model 2	0.923	141	5
Iris Model 3	0.953	137	3
Iris Model 4	0.714	126	23
Iris Model 5	0.832	131	12

It can be seen from Table 3.26 that Iris Model 3 recovered more of the true structure present in the data than the other models. This model also had only three observations where the model and the species classifications differed. Iris Model 2 also recovered most of the true structure present in the data. When considering these two models, Iris Model 3 would be the preferred model to be used as the assumption of equal covariance matrices for the three component distributions has greatly reduced the number of parameters to be estimated (Iris Model 3 requires 24 parameters to be estimated whilst Iris Model 2 requires the estimation of 44 parameters), and the model has recovered slightly more of the true structure present in the data than has Iris Model 2.

Iris Model 4 did not perform very well. This model only recovered 71.4% of the true structure present in the data set, and it had the lowest number of observations definitely assigned to a group. The program written for this thesis, in its

current version, does not allow for equal covariance matrices to be imposed for the component distributions, and it is suspected that this assumption would probably improve the extent of the structure being recovered. It would also reduce the number of parameters to be estimated under this partitioning from thirty two to twenty. Iris Model 4 also did not take into account any of the associations that existed between the variables.

The combination of the two categorical variables into one categorical variable having six levels (Iris Model 5), improved the performance of the model on all aspects investigated, with 83.2% of the structure being recovered and 131 of the observations being definitely assigned to a group. As well as performing better than Iris Model 4, Iris Model 5 also takes into account the associations that exist between the variables formed from categorising sepal width and sepal length. Once again, it is suspected that the assumption of equal covariance matrices for the component distributions would probably improve the agreement between the model and the species classification.

Of the two models fitted to the iris data set with sepal length and sepal width categorised, Iris Model 5 would be the more similar model to that of Everitt and Mérette, as account is taken in this model of the relationships known to exist between sepal length and sepal width. As pointed out in Section 1.8, Everitt and Mérette assumed that the underlying component distributions were the $N_4(\boldsymbol{\mu}_k, \Sigma)$ distribution for $k = 1, \ldots, 3$. They obtained an adjusted Rand index of 0.89 for their model, which assumed equal covariance matrices for the component distributions. We can conclude that the approach of multivariate finite mixture models has been fairly successful in clustering this data set.

In the next section, we use the approach of multivariate finite mixture models to cluster a data set which contains both categorical and continuous variables, and in which little is known *a priori* about the structure of the data.

3.4 Statistics Examination Data

This data was obtained from a statistics examination in which 66 individuals were assessed for their knowledge of the term's work. The examination was comprised of four long answer questions, and nine multiple choice questions, eight of which had five choices of answer, and one of which had four choices. The individuals were asked to attempt all questions. A copy of the examination paper is included in Appendix 2. The data set consists of the option selected for each of the nine multiple choice questions, and the mark given for each of the long answer questions. The multiple choice questions will be regarded as discrete variables, and the long answer marks as continuous variables.

A preliminary examination of the distributions of the continuous variables showed that three of the variables were approximately normally distributed. The fourth variable, Question 12, had a bimodal distribution with 25 of the observations receiving a mark of '0' and 18 individuals receiving a mark of '12'. For the initial analysis this variable will be considered to be normal. Initially a two component model will be fitted, and then the group structure will be investigated.

3.4.1 Two component models

We regard the data as a random sample from the distribution

$$f(\mathbf{x};\phi) = \sum_{k=1}^{2} \pi_k f_k(\mathbf{x};\theta_k)$$

where $\sum_{k=1}^{2} \pi_k = 1$, and $\pi_k \geq 0$ for $k = 1,2$; and $l = 1,13$. Under the model of complete independence for two clusters, the component distributions will be of the form

$$f_k(\mathbf{x}_i;\boldsymbol{\theta}_k) = \prod_{l=1}^{13} f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl}),$$

where $\boldsymbol{\theta}_{kl}$ is the parameter vector for group k, partition l, and $k = 1,2$; $l = 1,13$. For each of the 9 categorical variables, $f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl})$ is the $D(\lambda_{kl1},\ldots,\lambda_{klm_l})$ mass function[3], and for each of the 4 continuous variables, $f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl})$ is the $N(\mu_{kl},\sigma^2_{kl})$ density.

The notation that has previously been defined in section 3.2.1 will used. However, to distinguish between models previously used, 'Exam' will be included in the description of all models for the examination data. Hence, 'Exam Model i, n' denotes the model for the examination data that has partitioning i and n groups. The two group model of complete local independence will be referred to as Exam Model 1, 2.

This model was fitted iteratively using the EM algorithm with the initial grouping based on whether the student obtained a mark in Question 12 of ≤ 6 or > 6. Using this group assignment, the loglikelihood converged to -1372.057. The algorithm was also applied from a wide choice of starting values, and many other local maxima were found. The solution of the likelihood equation was taken to be the one corresponding to the largest of the local maxima, the same solution that was found using the grouping based on a Question 12 result of ≤ 6 or > 6. As previously, an observation $\mathbf{x}_i$ will be assigned to the population to which it has the highest estimated posterior probability of belonging.

---

[3] refer to notation in Chapter 2.

Table 3.27

Summary statistics for the 13 variables under Exam Model 1, 2.

(i) Categorical variables†

Question	Group	Level Probabilities				
1	1	0.170	**0.532**	0.191	0.085	0.021
	2	0.053	**0.632**	0.158	0.053	0.105
2	1	0.468	0.149	0.128	**0.255**	
	2	0.211	0.105	0.158	**0.526**	
3	1	0.000	0.064	**0.596**	0.149	0.191
	2	0.053	0.000	**0.895**	0.052	0.000
4	1	**0.511**	0.170	0.021	0.234	0.064
	2	**1.000**	0.000	0.000	0.000	0.000
5	1	0.064	0.042	**0.745**	0.085	0.064
	2	0.000	0.000	**0.895**	0.053	0.052
6	1	0.106	**0.702**	0.043	0.106	0.043
	2	0.000	**0.789**	0.000	0.158	0.053
7	1	0.043	**0.553**	0.277	0.042	0.085
	2	0.053	**0.737**	0.105	0.105	0.000
8	1	0.234	0.298	0.085	**0.213**	0.170
	2	0.158	0.263	0.053	**0.316**	0.210
9	1	0.469	**0.213**	0.043	0.191	0.085
	2	0.316	**0.421**	0.105	0.105	0.053

(ii) Continuous variables.

Question	Group	Mean	STDEV.
10	1	9.09	4.48
	2	11.42	3.96
11	1	7.75	4.45
	2	10.58	3.76
12	1	1.68	2.59
	2	11.95	0.22
13	1	4.72	4.58
	2	8.42	4.71

It can be seen from Table 3.27, that the continuous variables in Group 2 have a higher mean mark than those in Group 1. From the level probabilities, we can see that the individuals in Group 1 have a modal answer pattern '2, 1, 3, 1, 3, 2, 2, 2, 1',

† Figures in bold show the probability for correct answer

whereas the Group 2 individuals have a modal answer pattern of '2, 4, 3, 1, 3, 2, 2, 4, 2', the correct answer pattern. The Group 2 students appear to have done better in the examination than the Group 1 students. The estimated proportions in each group are 0.7121 and 0.2879, leading to expected numbers of 47.00 and 19.00 in each group.

An examination of the posterior probabilities showed that all 66 observations were 'definitely assigned' to a group, with posterior probabilities of assignment to that group of at least 0.999 for all observations.

The group assignment from Exam Model 1, 2 was used and the within group correlation structure was examined. No significant correlations were found for the four continuous variables. Tests of independence of the categorical variables were made for each group, and for all combinations of the categorical variables. The hypothesis of independence was rejected at the 0.05 level for Questions 4 and 5 in Group 1 only. When the multiple comparison effect is taken into account, it is possible that this hypothesis would not be rejected at the 0.05 level as it is expected that approximately two pairwise comparisons would be rejected in 36 comparisons due to chance alone. Since there would also be a large increase in the numbers of parameters to be estimated by combining these two variables, it was decided not to combine these two variables.

The composition of the group structure was then investigated. Further insight into the composition of the group structure may be gained by examining the mark the individuals received for the examination, and the students cumulative mark to date. The students cumulative mark for the year to date is comprised of a weighted contribution from the mark received for other items of assessment, together with a weighted contribution from the examination mark.

Table 3.28

Summary Statistics for Examination and Cumulative Marks under Exam Model 1, 2

	Group	Mean	STDEV
Examination Mark	1	40.40	13.12
	2	67.47	13.15
Cumulative Mark	1	52.13	13.21
	2	72.32	14.85

It can be seen from Table 3.28 that the Group 1 students have a lower examination mark and lower cumulative mark than do the students in Group 2. The group

stucture will be investigated by considering the examination marks for those who did not pass the examination, (i.e. an examination mark of less than 50%), and those who passed the examination (i.e. an examination mark of 50% or more). The cumulative mark will be also be investigated for each group by considering those students with a cumulative mark of less than 65% and those students with a cumulative mark of at least 65%.

Table 3.29

Examination Marks and Cumulative Marks under Exam Model 1, 2.

Group	Examination Mark		Cumulative Mark	
	$< 50\%$	$\geq 50\%$	$< 65\%$	$\geq 65\%$
1	37	10	39	8
2	2	17	4	15

It can be seen from Table 3.29 that that the majority (78.7%) of the students in Group 1 received an examination mark of less than 50%, whereas only 10.2% of the Group 2 students received an examination mark of less than 50%. It can also be seen that Group 2 students have a higher probability of having a cumulative grade of at least 65%, whilst the students in Group 1 more chance of having a cumulative grade of less than 65%. The Group 2 students definitely have higher statistics grades than do the Group 1 students.

When examining the statistics for the two groups found under Exam Model 1, 2 (refer to Table 3.27), it can be seen that in Group 2, for Question 12, the students have a mean mark of 11.95 and standard deviation of 0.045. The model appears to be grouping the observations on the basis of the marks received for Question 12. An investigation of the group structure and the mark received for Question 12 showed that Group 1 corresponded to those students who received a Question 12 mark of less than 11, whilst Group 2 was comprised of those students who received a Question 12 mark at least 11 (One student received a mark of 11 marks, and the remaining eighteen students in the group received 12 marks).

The influence of the students Question 12 mark on the grouping was also detected in the search for local maxima. When Exam Model 1, 2 was fitted from an initial classification based on whether the student obtained a mark of 12 for Question 12, the program 'crashed', as the program in its current version cannot cope computationally with a group having zero variance. There are ways to cope computationally with zero variance in a group. One method is to add 'random noise' to the variable concerned, and hence for example, a mark of 10 may be

represented as, 9.998 or 10.001 *etc.* Another method is to categorise the variable concerned.

Due to the bimodal nature of the Question 12 marks (33 students received a mark of no more than one mark, and 18 students received a mark of twelve marks), it was decided to categorise the Question 12 marks. Three different models were fitted to the data with the Question 12 marks categorised in different ways. In the first model, the Question 12 marks were categorised to three levels:- marks less than five, marks between five and eleven, and mark equal to 12. In the second model, the Question 12 marks were also categorised to three levels:- marks less than five, marks between five and ten, and mark of eleven or twelve. In the third model, the Question 12 marks were dichotomised:- mark less than eleven, and mark of eleven or twelve. Under the model of complete independence for the two clusters, each component distribution is now a product of ten discrete distributions and three univariate normal distributions. These three models all converged to solutions that gave the same group assignments as those under Exam Model 1, 2, and very similar estimates of the parameters for Questions 1 to 11 and Question 13 as those displayed in Table 3.27. All observations were assigned to a group with posterior probabilities of 1.00.

Due to the influence of the Question 12 marks on the group assignment, it was decided to categorise the Question 12 marks in a different way to see how this would affect the group assignment.

The marks received for Question 12 were categorised with three possible outcomes: mark ≤ 3, $4 \leq$ mark ≤ 8, and $9 \leq$ mark ≤ 12. A mixture of two component distributions was fitted. Under the model of complete independence for the two clusters, each component distribution is a product of ten discrete distributions and three univariate normal distributions. This model will be referred to as Exam Model 2, 2.

The model was fitted using the EM algorithm with the initial estimates of the classifications based on those from Exam Model 1, 2. The model converged to a solution where the loglikelihood was -1276.557. The algorithm was applied another eight times, where the initial classifications were generated by randomly splitting the observations into two groups. Six other solutions of the likelihood equation were found in the search for local maxima. The solution corresponding to the largest of the local maxima was used, the solution where the loglikelihood was -1273.299.

The classification of the observations to their class of greater probability under Exam Model 2, 2 is quite different to that given under Exam Model 1, 2. The

estimated proportions in each group are 0.3972 and 0.6028, leading to expected numbers in each group of 26.21 and 39.78.

Table 3.30

Summary statistics for the 13 variables under Exam Model 2, 2.

(i) Categorical variables†

Question	Group	Level Probabilities				
1	1	0.077	**0.429**	0.380	0.114	0.000
	2	0.175	**0.648**	0.051	0.051	0.075
2	1	0.533	0.077	0.232	**0.158**	
	2	0.302	0.175	0.074	**0.449**	
3	1	0.000	0.077	**0.455**	0.238	0.230
	2	0.025	0.025	**0.832**	0.044	0.074
4	1	**0.580**	0.191	0.000	0.115	0.114
	2	**0.699**	0.075	0.025	0.201	0.000
5	1	0.038	0.076	**0.771**	0.076	0.039
	2	0.050	0.000	**0.799**	0.076	0.075
6	1	0.114	**0.772**	0.038	0.038	0.038
	2	0.050	**0.698**	0.025	0.176	0.051
7	1	0.000	**0.587**	0.374	0.039	0.000
	2	0.075	**0.619**	0.130	0.075	0.101
8	1	0.376	0.271	0.114	**0.086**	0.153
	2	0.104	0.299	0.050	**0.346**	0.201
9	1	0.504	**0.190**	0.076	0.115	0.115
	2	0.372	**0.327**	0.050	0.201	0.050
12	1	0.699	0.192	0.109		
	2	0.494	0.075	0.431		

(ii) Continuous variables.

Question	Group	Mean	STDEV.
10	1	5.35	3.35
	2	12.67	2.11
11	1	7.27	4.54
	2	9.41	4.18
13	1	3.00	3.77
	2	7.63	4.71

It can be seen from Table 3.30, that the continuous variables in Group 2 have a higher mean mark than those in Group 1. Assuming within group independence of

† Figures in bold show the probability for correct answer

the questions, the answer pattern of highest probability for the discrete variables in Group 2 is '2, 4, 3, 1, 3, 2, 2, 4, 1, 1', and for Group 1 is '2, 1, 3, 1, 3, 2, 2, 1, 1, 1'. It can be seen for Question 9 that under Exam Model 2, 2 that option 1 is the choice of highest probability, whereas the correct anwer choice is option 2.

An examination of the posterior probabilities showed that 62 of the 66 observations are 'definitely assigned' to a group. The group assignment from Exam Model 2, 2 was used and the within group correlation structure was examined. No significant correlations were found for the three continuous variables. Tests of independence of the categorical variables were made for each group, and for all combinations of the categorical variables. The hypotheses of independence of the categorical variables were not rejected.

Analagously to Exam Model 1, 2, the group structure will be investigated by considering the mark students received for the examination and the students cumulative mark to date.

Table 3.31

Examination Marks and Cumulative Marks under Exam Model 2, 2.

Group	Examination Mark		Cumulative Mark	
	$< 50\%$	$\geq 50\%$	$< 65\%$	$\geq 65\%$
1	25	1	24	2
2	14	26	19	21

Table 3.31 indicates that most of the Group 1 students fail the examination and have a cumulative mark of less than 65%. With the Group 2 students, 65% of them pass the examination and 52.5% of them have a cumulative mark of at least 65%. It can be seen from a comparison of Table 3.29 and Table 3.31 that the composition of the groups has changed under the two models fitted.

An investigation of the differences between the group assignments under Exam Model 1 and Exam Model 2 showed that the classifications were basically nested. With the exception of the group assignments for three observations, Group 1 under Exam Model 2 is basically a subset of Group 1 under Exam Model 1, and Group 2 under Exam Model 1 is basically a subset of Group 2 under Exam Model 2. The change in the categorisation of the Question 12 marks for Exam Model 2 has definitely affected the group assignments. The group structure under the models fitted probably just reflects an underlying ordering of the observations with a different cut off point for the division of the observations into two groups. It is

possible that finite mixture analysis for this data set is an approximation to a continuous mixture.

So far we have assumed the existence of two groups in the data. It is of interest to see whether the likelihood ratio test of $H_0 : K = 1$ versus $H_a : K = 2$ would suggest that the data are a single population. The approximation[1] suggested by Wolfe (1971) was used and the likelihood ratio test was conducted. Under the assumption of a single population for the model, the algorithm converged to a solution where the loglikelihood was -1465.881. Under the assumption of a two group model, the loglikelihood is -1372.057. The test statistic $-2\log\lambda = 187.647$. The degrees of freedom for Wolfes approximation are 52. Hence we can be fairly confident that there is not a single population. It is not feasible to fit a model with three groups to this data due to the large increase in the parameters that would be required to be estimated from a small data set.

Exam Model 1, 2 is without question the preferred model for this data. It can be seen that the clusters detected using the approach of multivariate finite mixture models on this small data set are actually meaningful clusters, however further investigation would require a much larger data set. However, it should be noted that finite mixture analysis for this data set may possibly be an approximation to a continuous mixture.

We will conclude this chapter by giving a general stategy to be followed when using the approach of multivariate finite mixture models. The data should initially be examined for evidence of clusters by using some package such as Data Desk. Some exploratory data analysis should be carried out, seeking evidence of modality. But, as pointed out by Everitt and Hand (1981), unimodality of a distribution does not imply that the data is not a mixture.

Once it is decided that a mixture model is appropriate, the model of complete local independence should firstly be fitted. Then, using the group assignment from the model for complete independence, examine the within group correlation structure. Variables with strong within-cluster associations can be grouped together in a partition cell for the next series of fits. This process is repeated if necessary.

The number of groups to be fitted to the model should be checked by performing the likelihood ratio test. The posterior probabilities should also be examined. In all cases where the model is fitted, always check for the presence of local maxima. Finally, the clusters produced by the model should be examined to see if they are meaningful.

---

[1] Further details on this approximation may be found in section 1.5

Chapter 4

Models with missing Data

4.1 Introduction

Missing observations can often occur in sets of multivariate data. For example, in an archaeological study, often a specimen such as a skull may be damaged and some variables cannot be measured. Respondents in a survey may refuse to answer certain questions such as age or income. When plants or animals are the experimental units, they may die before all variables have been measured. In all of these cases, the data matrix may have rows in which not all variables have values. These missing data values may be scattered among the observations and have no particular pattern of occurrence.

Missing data values arising in this type of manner are not to be confused with the conceptual 'missing' or incomplete data formulation used in finite mixture model estimation, where each observation is regarded as 'missing' a label or a variable that indicates its component population of origin. The missing values described in the previous paragraph can be regarded as 'unintended' or 'accidental' missing data. Recent review papers in the literature on partially missing data include those by Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird and Rubin (1977) and Little (1982), and a monograph on partially missing data by Little and Rubin (1987).

A subsequent problem that occurs when some of the variables have not been observed for some of the observations, is selecting the appropriate method to use in the analysis of the data. Standard statistical methods have been developed to analyse rectangular data sets, where the rows of the data matrix represent the observations, and the columns represent the variables measured on each observation. When there are missing data values in a data set, several strategies are possible.

Statistical packages frequently exclude observations that have a missing value code for any of the variables involved in an analysis, and carry out a complete case analysis, using only the observations where the variables, $x_1, \ldots, x_p$, are all observed. This approach is generally easy to carry out, and may be satisfactory with small amounts of missing data. Any univariate statistics are then calculated on a common sample base of cases. However there is a potential loss of information by discarding the incomplete cases. If the number of missing values is large, it is

possible to lose considerable sample size. The critical concern however, is that this strategy can give biased estimates as it requires the strong assumption that the complete cases are a random subsample of the original observations. The completely recorded cases frequently differ from the original sample. For example, in a panel survey, observations that are lost to follow up often differ from those who remain in the study, and hence a complete case analysis can give quite biased results.

An alternative approach when missing data is present, is to consider all cases which have values for a particular variable. This procedure is known as available case analysis. Unlike complete cases analysis, there is no loss of potential information as all available values are used. However, this technique has the disadvantage that the sample base changes from variable to variable according to the pattern of missing data, and hence the statistics calculated can be based on different numbers of observations. The main disadvantage to this approach as pointed out by Everitt and Dunn (1991) for example, is that this procedure can also lead to covariance and correlation matrices that are not positive definite, which means that some linear combinations of variables will have zero or negative estimated variances.

Imputation procedures in which the missing values are filled in by some method such as mean imputation or regression imputation, are other procedures used to handle missing data. The resulting completed data set can then be analysed by standard statistical analyses. The performance of standard multivariate analyses with imputed data is unreliable, and it is hard to distinguish situations when the methods work from those when they fail. For example, Santos (1981) found that replacing the missing values with the sample means underestimates the variances and covariances. Dempster and Rubin (1983) point out that when imputation is used, it is easy to forget that the data is incomplete. Little and Rubin (1987) in chapters 2 to 4 demonstrate applications of imputation to designed experiments, multivariate analysis and sample surveys. Rubin (1991) discusses multiple imputation in which each missing value in a data set is replaced by a vector of m simulated values. Further references on imputation are given by Little and Rubin (1987), and Rubin (1991).

Procedures that are model based are another approach that can be used for multivariate incomplete data. A model for the data and a missing data mechanism are specified, and the parameters are estimated by some procedure such as maximum likelihood. This approach has advantages over other approaches. It is flexible and the model assumptions underlying the resulting methods can be evaluated. The log likelihood takes into account the incompleteness of the data and can be used to obtain large sample estimates of the variance from the second derivatives of the

log likelihood.

Knowledge of the mechanism that leads to the values being missing is also important in choosing the appropriate analysis to use and in interpretation of the results (Little and Rubin (1987)). This mechanism may be under the control of the researcher. For example, survey sampling could be viewed as leading to missing data: the desired survey variables are recorded for all individuals participating in the survey, and the variables are missing for the individuals not selected. If the sample is selected using a probability based sampling method, then the mechanism is under the control of the researcher and may be considered as ignorable. However, if some of the survey variables are missing for the sample, then the mechanism that leads to the observed data is not as well understood. An example in which the mechanism may not be under the control of the researcher but is understood, would be with censored data. In this instance, the data consists of the time to the occurrence of some event. At the end of the experiment, some of the data is censored as the event being monitored has not occurred prior to the completion of the experiment. For the censored data, we have partial information, namely that the failure time exceeds the time to censoring. Analysis of this type of data needs to take account of this type of information to avoid biased results.

Often the analysis of incomplete data proceeds with the assumption that the process that caused the missing data can be ignored (see Rubin (1976), and Little and Rubin (1987)). However, Little and Rubin (1987) in chapter 1, point out that the mechanism leading to missing data generally cannot be ignored as the performance of any method used in the analysis of multivariate data that has values recorded as missing, depends on the mechanism that gave rise to the missing data. Some methods perform adequately only when the missingness is not related to the observed variables. Other methods only require the weaker assumption where the probability of a variable being missing for a particular observation can depend on the observed variables for that observation but not on the values of the missing variables.

4.2 Likelihood based estimation for incomplete data

Maximum likelihood estimation for incomplete data is similar to that for complete data. The likelihood estimation of the parameters based on the incomplete data is derived, and the maximum likelihood estimates are found by solving the likelihood equation. However, Little and Rubin (1987) in chapter 5, point out that as the observed data are not generally an independent, identically distributed sample, the asymptotic standard errors obtained from the information matrix may not be relevant, and also the results that imply the large sample normality of the likelihood function do not immediately apply.

As previously, let $\mathbf{x}$ denote the data that would occur for each observation in the absence of missing values. We write $\mathbf{x}_i = (\mathbf{x}_{obs,i}, \mathbf{x}_{miss,i})$ where $\mathbf{x}_{obs,i}$ denotes the observed values for observation i, and $\mathbf{x}_{miss,i}$ denotes the missing values for observation i. This is a formal notation only and does not imply that the data are rearranged to achieve this pattern, the missing data values may be scattered among the observations and have no particular pattern of occurrence.

Rubin (1976) investigated the conditions under which the process that caused the missing data could be ignored, when making likelihood (and Bayesian) inferences about the parameters of the distribution of the data. He showed that the process that causes the missing data can be ignored when making likelihood based inferences about θ, the parameter of the data, if the data are 'missing at random', and the parameter, ϕ, of the missing data process is 'distinct' from θ. The correct likelihood is simply the density of the observed data, regarded as a function of the parameters.

'Missing at random' in the previous paragraph means that the probability of the observed pattern of missing and observed values may depend on the data that are observed, but given these observed values, not on the values of the missing data. That is, the distribution of the missing data mechanism does not depend on the missing values, $\mathbf{x}_{miss}$. In particular, the probability that a variable is missing for a particular observation can depend on the values of the observed variables for that subject, but not on the values of the missing variables. A formal definition of 'missing at random' is given by Rubin (1976).

Rubin (1976) defines ϕ to be 'distinct from θ' if the joint parameter space of (ϕ, θ) is the cartesian product of the parameter space of θ and the parameter space of ϕ. Hence if the data are missing at random, the missing data mechanism is ignorable in that the observed data likelihood is proportional to the complete data likelihood. That is,

$$L(\theta, \phi \mid \mathbf{x}_{obs}, R) \propto L(\theta \mid \mathbf{x}_{obs}),$$

where R is a missing data indicator with $R_{ij} = 1$ if x_{ij} is observed, and 0 if x_{ij} is missing. So inference based on the observed data likelihood is equivalent to inference based on the complete data likelihood.

For example, suppose that two variables, age, and income, are measured on n individuals. Suppose that the probability that income is recorded, varies according to the age of an individual, but does not vary according to the income of the individuals within an age group. Then the data are 'missing at random'. In this instance, the missingness is related to the observed data but not to the missing data.

A different approach to handling incomplete data is to treat the missing data as parameters and to maximise the complete data likelihood over the missing data and the parameters. That is, maximise the likelihood function

$$L_{miss}(\theta, \mathbf{x}_{miss} \mid \mathbf{x}_{obs}) = L(\theta \mid \mathbf{x}_{obs}, \mathbf{x}_{miss}) = f(\mathbf{x}_{obs}, \mathbf{x}_{miss} \mid \theta)$$

with respect to θ and $\mathbf{x}_{miss}$, where $f(\mathbf{x}_{obs}, \mathbf{x}_{miss} \mid \theta)$ is the density of the joint distribution of $\mathbf{x}_{obs}, \mathbf{x}_{miss}$ with parameter θ. Little and Rubin (1983) contend that the function L_{miss} is not a likelihood as the argument includes random variables which have a distribution under the model, and they therefore should not be treated as fixed parameters. Maximisation of L_{miss} with respect to θ and $\mathbf{x}_{miss}$ is not a maximum likelihood procedure from this perspective. Another problem with treating both θ and $\mathbf{x}_{miss}$ as parameters, is that the number of parameters increases as the number of observations increases. Little and Rubin (1983) maintain that this approach is not reliable, and show that this approach does not share the optimal properties of maximum likelihood estimation, except under the trivial asymptotics in which the proportion of missing data goes to zero as the sample size increases. See Little and Rubin (1987) for further references on the use of this procedure.

For some models and incomplete data patterns, the maximum likelihood estimates can be found by exploiting factorizations of the likelihood. Suppose we can find a parameterization, $\phi = \phi(\theta)$, where ϕ is a one to one monotone function of θ, and such that the log likelihood decomposes

$$\ell(\phi \mid \mathbf{x}_{obs}) = \ell_1(\phi_1 \mid \mathbf{x}_{obs}) + \ldots + \ell_J(\phi_J \mid \mathbf{x}_{obs}),$$

where (i) $\phi_1, \ldots, \phi_J$ are distinct parameters, and (ii) $\ell_j(\phi_j \mid \mathbf{x}_{obs})$ correspond to the loglikelihoods for complete or easier incomplete data problems, for $j = 1, \ldots, J$. Then, the loglikelihood can be maximised by maximising each component $\ell_j(\phi_j \mid \mathbf{x}_{obs})$, separately for each j. And if $\hat{\phi}$ is the maximum likelihood estimate of ϕ, then by the properties of maximum likelihood estimators, the maximum likelihood estimator of θ, $\hat{\theta}$ is $\hat{\theta} = \theta(\hat{\phi})$. Further information on factored likelihoods is given by Little and Rubin (1987). In practice, factorization of the likelihood is not usually possible.

We will now consider applying an iterative method of computation to obtain the maximum likelihood estimates when data are missing at random. Dempster, Laird and Rubin (1977) showed how the EM algorithm could be used to carry out the maximum likelihood estimation of incomplete data. Their general model includes both the conceptual missing data formulation used in finite mixture models and discussed in Chapter 1, and the unintended or accidental missing data values

described earlier in this chapter. For the latter case, Dempster, Laird and Rubin assumed that the missing values are missing at random, and that the parameters of the distribution of the data are independent of the parameters of the missing data process. They considered the multivariate normal model as one of their examples with missing data. Orchard and Woodbury (1972) presented a general approach to likelihood estimation for incomplete data. They described a cyclic algorithm for computing the maximum likelihood estimates of the parameters of the distribution when using incomplete data from a multivariate normal distribution. Beale and Little (1975) further developed the theory for multivariate normal distributions with missing data. Butler (1986) also considers the estimation of the parameters of a normal population with missing data.

Little and Rubin (1987) in Chapter 8, use the EM algorithm to compute the maximum likelihood estimates for the mean and covariance matrix for a random sample from the multivariate normal distribution where some of the data are missing at random. They give details of both the E step and the M step of the algorithm. The E step requires the use of the sweep operator, the details of which can be found in section 4.3. In the M step, the new estimates $\phi^{(t+1)}$ of the parameters are calculated from the estimated complete data sufficient statistics. It can be seen that the E step imputes the best linear predictors of the missing values, given the observed vaues and the current estimates of the parameters. It also calculates the adjustments to the covariance matrix needed to allow for imputation of the missing values.

Rubin and Szatrowski (1982) use the EM algorithm to find the maximum likelihood estimates for patterned covariance matrices. They found that the EM algorithm had the ability to handle simultaneously both missing data and patterned covariance matrices. We will also use the EM algorithm to simultaneously handle two types of problems.

We have previously used the EM algorithm for the computation of the likelihood estimates in the mixture problem by viewing the group assignment as the 'missing' data. We shall now use the EM algorithm to compute the likelihood estimates for mixture models, whilst simultaneously coping with both the missing data values in the data matrix and the 'missing' group assignments. We will initially consider, in Section 4.4, a mixture of bivariate normal distributions where data are missing at random. We will then extend the model to cope with p variables and in Section 4.5, we consider mixtures of multivariate normal distributions where data are missing at random. In section 4.6, we consider latent class analysis where data are missing at random, and in Section 4.7 we look at mixtures of location model variables with data missing at random. In section 4.8, we extend our ability to analyse data

using the multivariate mixture model by including the facility to handle data sets in which data are missing at random.

4.3 The Sweep Operator

In this section we describe the use of the SWEEP operator, originally defined by Beaton (1964). The sweep operator will be used when we consider the maximum likelihood estimation for the multivariate mixture model where continuous data are missing at random. The version of the sweep operator we describe is the one defined by Dempster (1969) and Goodnight (1979).

The sweep operator is defined for symmetric matrices as follows. A $p \times p$ symmetric matrix G is said to be swept on row and column k if it is replaced by another symmetric matrix H where

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk}, \qquad j \neq k \\ h_{jl} &= g_{jl} - g_{jk}g_{kl}/g_{kk}, \qquad j \neq k, l \neq k \\ &= g_{jl} - h_{jk}g_{kl}. \end{aligned} \tag{4.1}$$

We will use the notation SWP$[k]$G to denote the matrix defined by (4.1). The result of successively applying the operations SWP$[k_1]$, SWP$[k_2]$, ..., SWP$[k_t]$ to the matrix G will be denoted as SWP$[k_1, k_2, \ldots, k_t]$.

The sweep operator is commutative and associative, (Heiberger 1989) and more generally,

$$\text{SWP}[j_1, j_2, \ldots, j_t]\text{G} = \text{SWP}[k_1, k_2, \ldots, k_t]\text{G}$$

where $j_1, j_2, \ldots, j_t$ is any permutation of the set $k_1, k_2, \ldots, k_t$ (Little and Rubin 1987).

Suppose we have a sample of n observations on p variables, $X_1, \ldots, X_p$. Let G be the $(p+1) \times (p+1)$ matrix

$$G = \begin{pmatrix} 1 & \bar{X}_1 & \bar{X}_2 & \bar{X}_p \\ \bar{X}_1 & n^{-1}\Sigma X_1^2 & n^{-1}\Sigma X_1 X_2 & n^{-1}\Sigma X_1 X_p \\ \bar{X}_2 & n^{-1}\Sigma X_1 X_2 & n^{-1}\Sigma X_2^2 & n^{-1}\Sigma X_2 X_p \\ \vdots & \vdots & & \vdots \\ \bar{X}_p & n^{-1}\Sigma X_1 X_p & n^{-1}\Sigma X_2 X_p & n^{-1}\Sigma X_p^2 \end{pmatrix}$$

where $\bar{X}_1, \ldots, \bar{X}_p$ are the sample means, and the summations are over the p variables. If we index the rows and columns of the matrix G from 0 to p, row and

column j will then correspond to variable j. Sweeping on row and column 0 yields the matrix

$$\text{SWP[0]G} = \begin{pmatrix} -1 & \bar{X}_1 & \bar{X}_2 & & \bar{X}_p \\ \bar{X}_1 & s_{11} & s_{21} & & s_{p1} \\ \bar{X}_2 & s_{12} & s_{22} & & s_{p2} \\ \vdots & \vdots & & & \vdots \\ \bar{X}_p & s_{1p} & s_{2p} & & s_{pp} \end{pmatrix} \tag{4.2}$$

where s_{jk} is the sample covariance of X_j and X_k with factor n^{-1} instead of $(n-1)^{-1}$. Little and Rubin (1987) refer to (4.2) as the *augmented covariance matrix* of the variables $X_1, \ldots, X_p$. We can see that sweeping on row and column 0 corresponds to correcting the scaled cross products matrix of $X_1, \ldots, X_p$ for the means of $X_1, \ldots, X_p$ to create the covariance matrix. This process is also referred to as *sweeping on the constant term.*

Now suppose that SWP[0]G is swept on row and column 1. Then,

$$\begin{aligned} &\text{SWP}[0,1]\text{G} \\ &= \begin{pmatrix} -(1+\bar{x}_1^2/s_{11}) & \bar{x}_1/s_{11} & \bar{x}_2 - (s_{12}/s_{11})\bar{x}_1 & & \bar{x}_p - (s_{1p}/s_{11})\bar{x}_1 \\ & -1/s_{11} & s_{12}/s_{11} & & s_{1p}/s_{11} \\ & & s_{22} - s_{12}^2/s_{11} & & s_{2p} - s_{1p}s_{12}/s_{11} \\ & & & & \vdots \\ \bar{x}_p - (s_{1p}/s_{11})\bar{x}_1 & & & & s_{pp} - s_{1p}^2/s_{11} \end{pmatrix} \\ &= \begin{pmatrix} -A & B \\ B' & C \end{pmatrix} \end{aligned} \tag{4.3}$$

where A is a 2×2 matrix, B is the $2 \times (p-1)$ matrix in which the j^{th} column gives the intercept and the regression coefficients for the regression of $X_{(j+1)}$ on X_1, for $j = 1, \ldots, (p-1)$, and C is the $(p-1) \times (p-1)$ matrix containing the conditional covariance of $X_2, \ldots, X_p$ given X_1. For the case of $p = 2$, it can be seen from an examination of column 2 of (4.3), the familiar form of the conditional distribution (or regression) of X_2 on X_1.

The sweep operator is closely related to linear regression. Sweeping the constant term and the first q elements gives results for the multivariate regression of $X_{q+1}, \ldots, X_p$ on $X_1, \ldots, X_q$. Suppose that

$$\text{SWP}[0,1,\ldots,q]G = \begin{pmatrix} -D & E \\ E' & F \end{pmatrix},$$

then E is $(q+1) \times (p-q)$ matrix in which the j^{th} column of E gives the least squares intercept and the slopes of the regression of X_{j+q} on $X_1, \ldots, X_q$ for $j = 1, 2, \ldots, p-q$, F is the $(p-q) \times (p-q)$ conditional covariance matrix of $X_{q+1}, \ldots, X_p$ given $X_1, \ldots, X_q$, and D is a $(q+1) \times (q+1)$ matrix, the elements of which can be

used to give the variances and covariances of the estimated regression coefficients in E. Note that sweeping on a variable turns that variable from an outcome variable into a predictor variable. For further references on sweeping, see for example, Seber (1977).

The program written for this thesis to fit multivariate mixture models, sweeps on a matrix equivalent to that defined by (4.2). This matrix is created using the current estimates of the mean and covariance matrices.

We will now extend the approach of multivariate finite mixture models to include the problem of missing data. We will illustrate by firstly considering the maximum likelihood estimation for a mixture of incomplete bivariate normal distributions. We will assume that the data are missing at random. Little and Rubin (1987) on page 132, Example 7.3, demonstrate maximum likelihood estimation using the EM algorithm on bivariate normal data, where there is a general pattern of missing data on both variables. In the notation used for mixture models in this thesis, this example applies for the case of $K = 1$. In the following section, we will generalise this example to fit K groups in the mixture of bivariate normal distributions where data are missing at random. This relatively simple situation allows us to be fairly detailed without too much notational complexity.

4.4 Mixtures of bivariate normals

Suppose that we have two variables measured on n observations, and that data are missing at random throughout the data set. We regard the data, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, as a random sample from the distribution

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

where $\sum_{k=1}^{K} \pi_k = 1$, and $\pi_k \geq 0$, $k = 1, \ldots, K$. The component densities $f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)$ are the $N_2(\boldsymbol{\mu}_k, \Sigma_k)$ densities, with $\boldsymbol{\mu}_k = (\mu_{k1} \quad \mu_{k2})'$ and $\Sigma_k = \begin{pmatrix} \sigma_{k11} & \sigma_{k12} \\ \sigma_{k21} & \sigma_{k22} \end{pmatrix}$. Refer to section 2.2 for the discussion and estimates for mixtures of multivariate normals.

There are two types of missing data that have to be considered; one is the conceptual 'missing' data, the unobserved indicator of group membership, and the other is the unintended or accidental missing data values. In the 'complete data', we would know which group each observation came from, and the actual values of the missing variables. The hypothetical 'complete data', then, consists of the $n \times p$ data array that includes the observed data and the values of the missing data, 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 loglikelihood

$$\begin{aligned} L_C(\boldsymbol{\phi}) &= \log\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\left[\pi_k^{z_{ik}}\left\{f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\}^{z_{ik}}\right]\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{z_{ik}\log\pi_k + z_{ik}\log f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\pi_k - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left\{(\mathbf{x}_i-\boldsymbol{\mu}_k)'\,\Sigma_k^{-1}\,(\mathbf{x}_i-\boldsymbol{\mu}_k)\right. \\ &\qquad\qquad \left. + \log\left(2\pi\,|\Sigma_k|\right)\right\} \qquad (4.4) \end{aligned}$$

It can be seen that the complete data sufficient statistics are, for each class k, $\sum_{i=1}^{n} z_{ik}$, $\sum_{i=1}^{n} z_{ik}x_{i1}$, $\sum_{i=1}^{n} z_{ik}x_{i2}$, $\sum_{i=1}^{n} z_{ik}x_{i1}^2$, $\sum_{i=1}^{n} z_{ik}x_{i2}^2$, and $\sum_{i=1}^{n} z_{ik}x_{i1}x_{i2}$.

Assuming that the data are missing at random in the sense of Little and Rubin (1987), there are four cases that will have to be considered for a mixture of incomplete bivariate normal distributions:-

(a) Both x_{i1} and x_{i2} are observed.

(b) x_{i1} is observed, and x_{i2} is missing.

(c) x_{i1} is missing, and x_{i2} is observed.

(d) Both x_{i1} and x_{i2} are missing.

Case (d) is included as the estimates of the missing x_{i1} and x_{i2} may be required in the general model to be described in section 4.8.

At the t^{th} iteration , let $\theta_k^{(t)} = (\mu_k^{(t)}, \Sigma_k^{(t)})$ denote the current estimates of the parameters in group k. The E step of the algorithm requires the calculation of

$$Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)}) = E\{L_C(\boldsymbol{\phi}) \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}\},$$

the expectation of the complete data loglikelihood, conditional on the observed data and the current value of the parameters. We calculate $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)})$ as follows: It can be seen from (4.4) that each indicator variable z_{ik}, is replaced with

$$\begin{aligned} \hat{z}_{ik} &= E(z_{ik} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}) \\ &= P(z_{ik} = 1 \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_{obs,i}\,; \theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_{obs,i}\,; \theta_k^{(t)})} \end{aligned}$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that observation i belongs to group k.

Depending on the variables observed for observation i, the remaining expectations may require the use of the sweep operator described in section 4.3. This is adapted in the following manner:

Suppose that we form the augmented covariance matrix G, using the current estimates of the parameters for the k^{th} group.

$$G = \begin{pmatrix} -1 & \mu_{k1} & \mu_{k2} \\ \mu_{k1} & \sigma_{k11} & \sigma_{k12} \\ \mu_{k2} & \sigma_{k12} & \sigma_{k22} \end{pmatrix}$$

where, as in section 4.3, the rows and columns of G are indexed from 0 to 2. Then sweeping on row and column 1 corresponds to sweeping on x_{i1}, and sweeping on row and column 2 corresponds to sweeping on x_{i2}.

Suppose that for observation i, $\mathbf{x}_{obs,i} = x_{i1}$, that is, x_{i1} is observed and x_{i2} is missing. Sweeping on row and column 1 yields the matrix

$$SWP[1]G = \begin{pmatrix} -\left(1 + \frac{\mu_{k1}^2}{\sigma_{k11}}\right) & \frac{\mu_{k1}}{\sigma_{k11}} & \mu_{k2} - \frac{\sigma_{k12}}{\sigma_{k11}}\mu_{k1} \\ \frac{\mu_{k1}}{\sigma_{k11}} & \frac{-1}{\sigma_{k11}} & \frac{\sigma_{k12}}{\sigma_{k11}} \\ \mu_{k2} - \frac{\sigma_{k12}}{\sigma_{k11}}\mu_{k1} & \frac{\sigma_{k12}}{\sigma_{k11}} & \sigma_{k22} - \frac{\sigma_{k12}^2}{\sigma_{k11}} \end{pmatrix} \tag{4.5}$$

It can be seen from (4.5), that row and column 2 give the slope, $(\sigma_{k12}/\sigma_{k11})$, the intercept, $(\mu_{k2} - \mu_{k1}\sigma_{k12}/\sigma_{k11})$, and the residual variance, $(\sigma_{k22} - \sigma_{k12}^2/\sigma_{k11})$ of the regression of x_{i2} on x_{i1}, for group k. From the properties of the bivariate normal distribution, see for example Bain and Englehardt (1992), we know that the conditional distribution of $x_{i2} \mid x_{i1}$ is normal with mean $\mu_2 + (x_{i1} - \mu_1)\sigma_{12}/\sigma_{11}$ and variance $\sigma_{22} - \sigma_{12}^2/\sigma_{11}$. Thus, if $\mathbf{x}_{obs,i} = x_{i1}$, the conditional means and covariances if observation i is in group k, are found from the current parameter estimates for group k by the sweep operator.

The remaining expectations to be calculated in the E step are as follows.

$$\begin{aligned} &E(z_{ik}x_{i1} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) \\ &\quad = E\left(z_{ik} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)}\right) E\left(x_{i1} \mid \mathbf{x}_{obs,i}\,; \theta_k^{(t)}\right) \\ &\quad = \begin{cases} \hat{z}_{ik}x_{i1} & x_{i1} \text{ observed,} \\ \hat{z}_{ik}\left(\mu_{k1} + \frac{\sigma_{k12}}{\sigma_{k22}}(x_{i2} - \mu_{k2})\right) & x_{i1} \text{ missing, } x_{i2} \text{ observed,} \\ \hat{z}_{ik}\mu_{k1} & \text{both } x_{i1} \text{ and } x_{i2} \text{ missing.} \end{cases} \end{aligned}$$

$$E(z_{ik}x_{i2} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)})$$

$$= \begin{cases} \hat{z}_{ik}x_{i2} & x_{i2} \text{ observed,} \\ \hat{z}_{ik}\left(\mu_{k2} + \dfrac{\sigma_{k12}}{\sigma_{k11}}(x_{i1} - \mu_{k1})\right) & x_{i1} \text{ observed, } x_{i2} \text{ missing,} \\ \hat{z}_{ik}\mu_{k2} & \text{both } x_{i1} \text{ and } x_{i2} \text{ missing.} \end{cases}$$

$$E(z_{ik}x_{i1}x_{i2} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)})$$

$$= \begin{cases} \hat{z}_{ik}x_{i1}x_{i2} & \text{both } x_{i1} \text{ and } x_{i2} \text{ observed,} \\ \hat{z}_{ik}x_{i1}\left(\mu_{k2} + \dfrac{\sigma_{k12}}{\sigma_{k11}}(x_{i1} - \mu_{k1})\right) & x_{i1} \text{ observed, } x_{i2} \text{ missing,} \\ \hat{z}_{ik}x_{i2}\left(\mu_{k1} + \dfrac{\sigma_{k12}}{\sigma_{k22}}(x_{i2} - \mu_{k2})\right) & x_{i1} \text{ missing, } x_{i2} \text{ observed,} \\ \hat{z}_{ik}\sigma_{k12} & \text{both } x_{i1} \text{ and } x_{i2} \text{ missing.} \end{cases}$$

$$E(z_{ik}x_{i1}^2 \mid \mathbf{x}_{obs,i}, \theta_k^{(t)})$$

$$= E\left(z_{ik} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)}\right) E\left(x_{i1}^2 \mid \mathbf{x}_{obs,i}\,; \theta_k^{(t)}\right)$$

$$= \begin{cases} \hat{z}_{ik}x_{i1}^2 & x_{i1} \text{ observed,} \\ \hat{z}_{ik}\left[\left(E\left(x_{i1} \mid \mathbf{x}_{obs,i}\,; \theta_k^{(t)}\right)\right)^2 + \text{ Var } \left(x_{i1} \mid \mathbf{x}_{obs,i}\,; \theta_k^{(t)}\right)\right] & x_{i1} \text{ missing,} \end{cases}$$

$$= \begin{cases} \hat{z}_{ik}x_{i1}^2 & x_{i1} \text{ observed,} \\ \hat{z}_{ik}\left(\left(\mu_{k1} + \dfrac{\sigma_{k12}}{\sigma_{k22}}(x_{i2} - \mu_{k2})\right)^2 + \sigma_{k11} - \dfrac{\sigma_{k12}^2}{\sigma_{k22}}\right) & x_{i1} \text{ missing, } x_{i2} \text{ observed,} \\ \hat{z}_{ik}\left(\sigma_{k11} + \mu_{k1}^2\right) & \text{both } x_{i1} \text{ and } x_{i2} \text{ missing.} \end{cases}$$

$$E(z_{ik}x_{i2}^2 \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) \quad =$$

$$\begin{cases} \hat{z}_{ik}x_{i2}^2 & x_{i2} \text{ observed,} \\ \hat{z}_{ik}\left(\left(\mu_{k2} + \dfrac{\sigma_{k12}}{\sigma_{k11}}(x_{i1} - \mu_{k1})\right)^2 + \sigma_{k22} - \dfrac{\sigma_{k12}^2}{\sigma_{k11}}\right) & x_{i1} \text{ observed, } x_{i2} \text{ missing,} \\ \hat{z}_{ik}\left(\sigma_{k22} + \mu_{k2}^2\right) & \text{both } x_{i1} \text{ and } x_{i2} \text{ missing.} \end{cases}$$

The missing x_{ij}, $[x_{ij}^2]$, are replaced by the conditional mean of x_{ij}, $[x_{ij}^2]$, given the set of values $x_{obs,i}$ observed for that observation. These conditional means and the nonzero conditional covariances are found from the current parameter estimates by sweeping the augmented covariance matrix (refer to equation 4.2) on the observed

variable. This gives the results for the regression of x_{i1} on x_{i2} if x_{i2} is missing and the regression of x_{i2} on x_{i1} if x_{i1} is missing.

At the M-step, $\phi^{(t+1)}$ is chosen to be the value of ϕ that maximises $Q(\phi, \phi^{(t)})$ with respect to ϕ.

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ik}$$

$$\hat{\mu}_{kj}^{(t+1)} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \hat{x}_{ij,k}^{(t)}$$

$$\hat{\sigma}_{kjj'}^{(t+1)} = \frac{1}{n\hat{\pi}_k} E\left(\sum_{i=1}^{n} z_{ik} x_{ij} x_{ij'} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}\right) - \hat{\mu}_{kj}^{(t+1)} \hat{\mu}_{kj'}^{(t+1)}$$

for $j, j' = 1, 2$ and $k = 1, \ldots, K$, and where

$$\hat{x}_{ij,k}^{(t)} = \begin{cases} x_{ij} & \text{if } x_{ij} \text{ is observed,} \\ E(x_{ij} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) & \text{if } x_{ij} \text{ is missing.} \end{cases}$$

The EM algorithm alternates between the E step and the M step until convergence.

In the next section, we will consider the maximum likelihood estimation of a mixture of incomplete multivariate normal distributions, where data are missing at random.

4.5 Mixtures of multivariate normals

Suppose that p attributes are measured on n individuals. We regard the data, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, as a random sample from the distribution

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

where $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$, for $k = 1, \ldots, K$.

Under the normality assumption, the component densities are the $N_p(\boldsymbol{\mu}_k, \Sigma_k)$ densities. Refer to section 2.2 for the estimates for mixtures of multivariate normals where there are no missing data values.

Suppose that some of the data are missing at random, scattered throughout the data set. The observation vector $\mathbf{x}_i$, is written in the form $(\mathbf{x}_{obs,i}, \mathbf{x}_{miss,i})$, as described in section 4.2. In fitting the mixture model, there are two types of missing data that have to be considered; one is the conceptual 'missing' data, the

unobserved indicator of group membership, and the other is the unintended or accidental missing data values.

The EM algorithm is applied to the mixture model where the indicator variables $\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_n$ are as described in section 2.2. The hypothetical 'complete data', then, consists of the $n \times p$ data array that includes the observed data and the values of the missing data, and the conceptual $n \times K$ array of class membership indicators.

The loglikelihood for the hypothetical 'complete data' is

$$\begin{aligned} L_C(\boldsymbol{\phi}) &= \log\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\left[\pi_k^{z_{ik}}\left\{f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\}^{z_{ik}}\right]\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{z_{ik}\log\pi_k + z_{ik}\log f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\pi_k - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left\{(\mathbf{x}_i-\boldsymbol{\mu}_k)'\,\Sigma_k^{-1}\,(\mathbf{x}_i-\boldsymbol{\mu}_k)\right. \\ &\qquad\qquad \left. + \log\left(2\pi\,|\Sigma_k|\right)\right\} \end{aligned}$$

The complete data sufficient statistics for the model are

(*i*) $\sum_{i=1}^{n} z_{ik}$, for each group k;

(*ii*) $\sum_{i=1}^{n} z_{ik}x_{ij}$ for each group k, and each (continuous) variable x_{ij};

(*iii*) $\sum_{i=1}^{n} z_{ik}x_{ij}x_{ij'}$ for each group k, and each pair of (continuous) variables, x_{ij} and $x_{ij'}$.

The EM algorithm alternates between the two calculations, the E step and the M step until convergence. At the t^{th} iteration , let $\theta_k^{(t)} = (\mu_k^{(t)}, \Sigma_k^{(t)})$ denote the current estimates of the parameters for group k. The E step of the EM algorithm requires the calculation of

$$Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)}) = E\{L_C(\boldsymbol{\phi}) \mid \mathbf{x}_{obs}; \boldsymbol{\phi}^{(t)}\},$$

the expectation of the complete data loglikelihood, conditional on the observed data and the current value of the parameters. We calculate $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)})$ by replacing z_{ik} with

$$\begin{aligned} \hat{z}_{ik} = \hat{z}_{ik}^{(t)} &= E(z_{ik} \mid \mathbf{x}_{obs,i}\,; \boldsymbol{\phi}^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_{obs,i}\,; \theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_{obs,i}\,; \theta_k^{(t)})} \end{aligned}$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that observation i belongs to group k. When t is not clear from the context, we use the full

form $\hat{z}_{ik}^{(t)}$. Depending on the variables observed for observation i, the remaining expectations to be calculated in the E step, may require the use of the sweep operator described in section 4.3.

$$E(z_{ik}x_{ij} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)}) = \begin{cases} \hat{z}_{ik}x_{ij} & x_{ij} \text{ present,} \\ \hat{z}_{ik}E(x_{ij} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)}) & x_{ij} \text{ missing.} \end{cases}$$

$$\begin{aligned} &E(z_{ik}x_{ij}^2 \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) \\ &\quad = E\left(z_{ik} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)}\right) E\left(x_{ij}^2 \mid \mathbf{x}_{obs,i}\,; \theta_k^{(t)}\right) \\ &\quad = \begin{cases} \hat{z}_{ik}x_{ij}^2 & x_{ij} \text{ observed,} \\ \hat{z}_{ik}\left[\left(E\left(x_{ij} \mid \mathbf{x}_{obs,i}\,; \theta_k^{(t)}\right)\right)^2 + \mathrm{Var}\left(x_{ij} \mid \mathbf{x}_{obs,i}\,; \theta_k^{(t)}\right)\right] & x_{ij} \text{ missing.} \end{cases} \end{aligned}$$

For $j \neq j'$,

$$\begin{aligned} &E(z_{ik}x_{ij}x_{ij'} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)}) \\ &\quad = \begin{cases} \hat{z}_{ik}x_{ij}x_{ij'} & \text{both } x_{ij} \text{ and } x_{ij'} \text{ present,} \\ \hat{z}_{ik}x_{ij}E(x_{ij'} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)}) & x_{ij} \text{ present, } x_{ij'} \text{ missing,} \\ \hat{z}_{ik}E(x_{ij} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)})x_{ij'} & x_{ij} \text{ missing, } x_{ij'} \text{ present,} \\ \hat{z}_{ik}\left[E(x_{ij} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)})\, E(x_{ij'} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)})\right. & \\ \qquad\qquad \left. + \mathrm{Cov}\left(x_{ij},\, x_{ij'} \mid \mathbf{x}_{obs,i}\,; \theta_k^{(t)}\right)\right] & \text{both } x_{ij} \text{ and } x_{ij'} \text{ missing.} \end{cases} \end{aligned}$$

It can be seen from the above expectations, that when there is only one factor, x_{ij}, missing, the missing x_{ij}, are replaced by the conditional mean of x_{ij}, given the set of values, $\mathbf{x}_{obs,i}$, observed for that observation and the current estimates of the parameters. However, for the conditional expectations to be used in the calculation of the variance covariance matrix, *i.e.* $E(z_{ik}x_{ij}^2 \mid \mathbf{x}_{obs,i}; \theta_k^{(t)})$ and $E(z_{ik}x_{ij}x_{ij'} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)})$, then respectively if x_{ij} is missing, or if both x_{ij} and $x_{ij'}$ are missing, the conditional mean of x_{ij} is adjusted by the conditional covariances as shown above. These conditional means and the nonzero conditional covariances are found by using the sweep operator described in section 4.3 on the augmented covariance matrix that is created using the current estimates of the parameters. The augmented covariance matrix is swept on the observed variables $\mathbf{x}_{obs,i}$, such that these variables are the predictors in the regression equation and the remaining variables are the outcome variables.

In the M step of the algorithm, the new estimates $\theta^{(t+1)}$ of the parameters are estimated from the complete data sufficient statistics.

$$\hat{\pi}_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ik}^{(t)}$$

$$\hat{\mu}_{kj}^{(t+1)} = \frac{1}{n\hat{\pi}_k} E\left(\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_{ij} \mid \mathbf{x}_{obs,i},\, \theta_k^{(t)}\right)$$

$$\hat{\Sigma}_{kjj'}^{(t+1)} = \frac{1}{n\hat{\pi}_k} E\left(\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_{ij} x_{ij'} \mid \mathbf{x}_{obs,i},\, \theta_k^{(t)}\right) - \hat{\mu}_{kj}^{(t+1)} \hat{\mu}_{kj'}^{(t+1)}$$

Because of the adjustment required to be made to the conditional means when both x_{ij} and $x_{ij'}$ are missing, it is convenient to use similar notation to that used by Little and Rubin (1987, page 144). The conditional covariance between variables j and j' for observation i, given that in group k, is defined as

$$C_{ki,jj'}^{(t)} = \begin{cases} 0 & \text{if either } x_{ij} \text{ or } x_{ij'} \text{ are observed,} \\ \text{Cov}\,(x_{ij}, x_{ij'} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) & \text{if both } x_{ij} \text{ and } x_{ij'} \text{ are missing,} \end{cases}$$

and the imputed value for variable j of observation i, given the current value of the parameters and that observation i is in group k, is defined as

$$\hat{x}_{ij,k}^{(t)} = \begin{cases} x_{ij} & \text{if } x_{ij} \text{ is present,} \\ E(x_{ij} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) & \text{if } x_{ij} \text{ is missing.} \end{cases}$$

The parameter estimates for the mean and the variance covariances can thus be written in the form

$$\begin{aligned} \hat{\mu}_{kj}^{(t+1)} &= \frac{1}{n\hat{\pi}_k} E\left(\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_{ij} \mid \mathbf{x}_{obs,i},\, \theta_k^{(t)}\right) \\ &= \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \hat{x}_{ij,k}^{(t)} \end{aligned}$$

$$\begin{aligned} \hat{\Sigma}_{kjj'}^{(t+1)} &= \frac{1}{n\hat{\pi}_k} E\left(\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_{ij} x_{ij'} \mid \mathbf{x}_{obs,i},\, \theta_k^{(t)}\right) - \hat{\mu}_{kj}^{(t+1)} \hat{\mu}_{kj'}^{(t+1)} \\ &= \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[(\hat{x}_{ij,k}^{(t)} - \hat{\mu}_{kj}^{(t+1)})(\hat{x}_{ij',k}^{(t)} - \hat{\mu}_{kj'}^{(t+1)}) + C_{ki,jj'}^{(t)}\right] \end{aligned}$$

for $j, j' = 1, \ldots, p$ and $k = 1, \ldots, K$. The EM algorithm alternates between the two calculations, the E step and the M step until convergence.

In the next section, we consider latent class analysis where we have a general pattern of missing data.

4.6 Latent class analysis with data missing at random

Suppose that p attributes are measured on n individuals, and that some of the data are missing at random. We regard the data, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, as a random sample from the distribution

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

where $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$, for $k = 1, \ldots, K$.

Let $\mathbf{x} = (x_1, \ldots, x_p)'$ be the response vector observed on each observation, where the jth variable can have levels numbered from 1 to M_j. Let λ_{kjm} be the probability that variable j takes level m in group k.

Assuming independence of the variables within each of the K sub-populations, the component distributions are of the form

$$f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) = \prod_{j=1}^{p} \lambda_{kjx_{ij}}$$

where $\boldsymbol{\theta}_k$, the parameter vector for distribution f_k, are the $\{\lambda_{kjm}\}$. As described in section 2.3, $\sum_{m=1}^{M_j} \lambda_{kjm} = 1$ for any fixed k and j.

Now suppose that for each variable j for $j = 1, \ldots, p$, we define an indicator variable

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

Then, the density function for observation i in group k can be written as

$$f_k(\mathbf{x}_i; \theta_k) = \prod_{j=1}^{p} \prod_{m=1}^{M_j} \lambda_{kjm}^{\delta_{ijm}}.$$

The EM algorithm is applied to the mixture model. Let the indicators of group membership $\mathbf{z}_1, \ldots, \mathbf{z}_n$, be as defined in section 2.2. With a general pattern of missing data, there are two types of missing data that have to be considered; one is the conceptual 'missing' indicator of group membership, and the other is the unintended or accidental missing data values. The 'complete data' thus consists of the $n \times p$ data array that includes the observed data and the values of the missing data, and the conceptual $n \times K$ array $\{z_{ik}\}$ of class membership indicators.

The likelihood for the complete data set can be written as

$$\begin{aligned} \ell_C(\boldsymbol{\phi}) &= \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}} f_k(\mathbf{x}; \theta_k)^{z_{ik}} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}} \left\{ \prod_{j=1}^{p} \prod_{m=1}^{M_j} \lambda_{kjm}^{\delta_{ijm}} \right\}^{z_{ik}} \end{aligned}$$

The loglikelihood for the hypothetical complete data is

$$L_C(\boldsymbol{\phi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left(z_{ik} \left\{ \log \pi_k + \sum_{j=1}^{p} \sum_{m=1}^{M_j} \delta_{ijm} \log(\lambda_{kjm}) \right\} \right).$$

The complete data sufficient statistics for the model are,

(*i*) $\sum_{i=1}^{n} z_{ik}$ for each group k, and

(*ii*) $\sum_{i=1}^{n} z_{ik}\delta_{ijm}$ for each group k, each categorical variable x_j, and each value m of x_j.

The E step of the EM algorithm requires the calculation of

$$Q(\phi, \phi^{(t)}) = E\{L_C(\phi) \mid \mathbf{x}_{obs}; \phi^{(t)}\},$$

the expectation of the complete data loglikelihood, conditional on the observed data and the current value of the parameters. We calculate $Q(\phi, \phi^{(t)})$ as follows. We replace z_{ik} with

$$\begin{aligned}\hat{z}_{ik} &= E(z_{ik} \mid \mathbf{x}_{obs,i}\,; \phi^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_{obs,i}\,; \theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_{obs,i}\,; \theta_k^{(t)})}\end{aligned}$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that observation i belongs to group k. The E step also calculates

$$\begin{aligned}E(z_{ik}\delta_{ijm} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) &= \begin{cases} \hat{z}_{ik}\delta_{ijm} & \text{if } x_{ij} \text{ is observed} \\ \hat{z}_{ik}E(\delta_{ijm} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) & \text{if } x_{ij} \text{ is not observed} \end{cases} \\ &= \begin{cases} \hat{z}_{ik}\delta_{ijm} & \text{if } x_{ij} \text{ is observed,} \\ \hat{z}_{ik}\lambda_{kjm}^{(t)} & \text{if } x_{ij} \text{ is not observed.} \end{cases}\end{aligned}$$

The missing indicators, δ_{ijm}, are replaced by the current value of the $\lambda_{kjm}^{(t)}$. Let

$$\hat{\delta}_{ijm} = \begin{cases} \delta_{ijm} & \text{if } x_{ij} \text{ is observed} \\ \lambda_{kjm} & \text{if } x_{ij} \text{ is not observed.} \end{cases}$$

Then, we can write the expectation in the form

$$E(z_{ik}\delta_{ijm} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) = \hat{z}_{ik}\hat{\delta}_{ijm}$$

At the M step, $\phi^{(t+1)}$ is chosen to be the value of ϕ that maximises $Q(\phi, \phi^{(t)})$ with respect to ϕ.

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ik}$$

$$\hat{\lambda}_{kjm} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n} \hat{z}_{ik}\hat{\delta}_{ijm}$$

for $k = 1, \ldots, K$, $j = 1, \ldots, p$ and $m = 1, \ldots, M_j$.

The EM algorithm alternates between the E step and the M step until convergence.

In section 4.5 we considered mixture models in which the component distributions were multivariate normal, and data were missing at random throughout the data set. In section 4.6, mixture models were considered for categorical variables with missing data values. In the next section we discuss missing data methods for mixture models in which the component distributions are location model distributions.

Little and Schlucter (1985) and Little and Rubin (1987) in chapter 10 discuss the location model with missing data values. They demonstrate the use of the EM algorithm for the maximum likelihood estimation of the parameters given the observed data. In the notation used in this thesis, this example applies for the fitting of a mixture of location models for the case $k = 1$. In the next section, this is extended to apply for a mixture of K groups.

4.7 Mixtures of location models

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$, be a random sample from the finite mixture distribution

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

where $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$, for $k = 1, \ldots, K$.

Suppose that $p + 1$ attributes are measured on the n individuals, where the vector of variables, $\mathbf{x} = (x_1, \ldots, x_{p+1})'$, has p continuous variables, and one discrete variable. The component distributions, $f_k(\mathbf{x}; \boldsymbol{\theta}_k)$ have the following form. The discrete variable takes the values $1, \ldots, M$ with probabilities $\lambda_{k1}, \ldots, \lambda_{kM}$, where $\sum_{m=1}^{M} \lambda_{km} = 1$. Conditional on the discrete variable taking value m, the p continuous variables have the multivariate normal distribution, $N_p(\boldsymbol{\nu}_{km}, \Xi_k)$. That is, each component distribution is the location model of Olkin and Tate (1961). Refer to section 2.4 for discussions on mixtures of location models.

To distinguish between the categorical and continuous variables, in this section, u will be used to denote the discrete variable and $\mathbf{v}$ will be used to denote the $p \times 1$ vector of continuous variables. Thus, $\mathbf{x}$ takes the form $(u, \mathbf{v})$.

Suppose that some of the data are missing at random, and are scattered throughout the data set. These unintended missing values can be either continuous or categorical variables. The EM algorithm is applied to the mixture model, where the indicators of group membership $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are as described in section 2.2. The 'complete data' consists of the $n \times p$ data array that includes the observed data

and the values of the missing data, and the conceptual $n \times K$ array $\{z_{ik}\}$ of class membership indicators.

The loglikelihood for the hypothetical 'complete data' is

$$\begin{aligned} L_C(\phi) &= \log\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\left[\pi_k^{z_{ik}}\left\{f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\}^{z_{ik}}\right]\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{z_{ik}\log\pi_k + z_{ik}\log f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\left[z_{ik}\log\pi_k + z_{ik}\left\{\sum_{m=1}^{M} w_{im}\log\lambda_{km} + h(\Xi_k)\right.\right. \\ &\qquad \left.\left. -\frac{1}{2}\left(\mathbf{v}_i - \sum_{m=1}^{M} w_{im}\boldsymbol{\nu}_{km}\right)' \Xi_k^{-1}\left(\mathbf{v}_i - \sum_{m=1}^{M} w_{im}\boldsymbol{\nu}_{km}\right)\right\}\right] \end{aligned}$$

where $h(\Xi_k) = \frac{1}{2}np\log(2\pi) + \log(|\Xi_k|)$, and $w_{im} = \begin{cases} 1 & \text{if } u_i = m; \\ 0 & \text{if } u_i \neq m. \end{cases}$

The complete data sufficient statistics for the model are

(*i*) $\sum_{i=1}^{n} z_{ik}$, for each group k;

(*ii*) $\sum_{i=1}^{n} z_{ik} w_{im} v_{ij}$ for each class k, each continuous variable v_j and each value m of the categorical variable u;

(*iii*) $\sum_{i=1}^{n} z_{ik} v_{ij} v_{ij'}$, for each class k, each pair of continuous variables v_j and $v_{j'}$, for $j \leq j'$.

Note that all this is the same as in section 2.4, but now u_i and /or some of the v_{ij} are missing.

The EM algorithm alternates between the two calculations, the E step and the M step. At the t^{th} iteration, let $\theta_k^{(t)} = (\lambda_k^{(t)}, \nu_{mk}^{(t)}, \Xi_k^{(t)}\,;\, m = 1, \ldots, M)$ denote the current estimates of the parameters for group k. The E step of the algorithm requires the calculation of $Q(\phi, \phi^{(t)})$, the expectation of the complete data loglikelihood, conditional on the observed data and the current value of the parameters.

We calculate $Q(\phi, \phi^{(t)})$ by replacing z_{ik} with

$$\begin{aligned} \hat{z}_{ik} &= E(z_{ik} \mid \mathbf{x}_{obs,i}\,;\phi^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_{obs,i}\,;\theta_k^{(t)})}{\sum_{k=1}^{K}\pi_k f_k(\mathbf{x}_{obs,i}\,;\theta_k^{(t)})} \end{aligned}$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that observation i belongs to group k. Depending on the variables observed for observation

i, the remaining expectations may require the use of the sweep operator described in section 4.3. These expectations are calculated as follows.

$$\begin{aligned} E(z_{ik}w_{im} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}) &= E(z_{ik} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}) E_{P_k}(w_{im} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}) \\ &= \hat{z}_{ik}^{(t)} P\left[w_{im} = 1 \mid \mathbf{x}_{obs,i}\,;\, \left\{\lambda_k^{(t)}, \nu_{mk}^{(t)}, \Xi_k^{(t)}\right\}\right] \end{aligned}$$

where $E_{P_k}(w_{im} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)})$ is the expectation of w_{im} given $\mathbf{x}_{obs,i}$, and P_k is the k^{th} population in the mixture with parameters taken to be those at the t^{th} iteration.

This expectation involves the conditional posterior probability that observation i has u_i at level m given the observed continuous variables, and the current values of the parameters. All parameters that are in the expressions that follow, are equal to the current parameter estimates in $\theta^{(t)}$.

If the categorical variable is not observed, then

$$\begin{aligned} & P\left[w_{im} = 1 \mid \mathbf{x}_{obs,i}\,;\, \{\lambda_k, \nu_{km}, \Xi_k\}\right] \\ &= \frac{f\left[\mathbf{x}_{obs,i}\,;\, \{\lambda_k, \nu_{km}, \Xi_k\} \mid w_{im} = 1\right] \lambda_{km}}{\sum_{m=1}^{M} f\left[\mathbf{x}_{obs,i}\,;\, \{\lambda_k, \nu_{km}, \Xi_k\} \mid w_{im} = 1\right] \lambda_{km}} \\ &= \frac{\lambda_{km} \exp\left(-\frac{1}{2}\left(\mathbf{x}_{obs,i} - \boldsymbol{\nu}_{obs,i,km}\right)' \Xi_{k,obs,i}^{-1} \left(\mathbf{x}_{obs,i} - \boldsymbol{\nu}_{obs,i,km}\right)\right)}{\sum_{m=1}^{M} \lambda_{km} \exp\left(-\frac{1}{2}\left(\mathbf{x}_{obs,i} - \boldsymbol{\nu}_{obs,i,km}\right)' \Xi_{k,obs,i}^{-1} \left(\mathbf{x}_{obs,i} - \boldsymbol{\nu}_{obs,i,km}\right)\right)} \\ &= \frac{\lambda_{km} \exp\left(\mathbf{x}_{obs,i} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km} - \frac{1}{2} \boldsymbol{\nu}'_{obs,i,km} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km}\right)}{\sum_{m=1}^{M} \lambda_{km} \exp\left(\mathbf{x}_{obs,i} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km} - \frac{1}{2} \boldsymbol{\nu}'_{obs,i,km} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km}\right)} \\ &= \frac{\exp\left(\mathbf{x}_{obs,i} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km} - \frac{1}{2} \boldsymbol{\nu}'_{obs,i,km} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km} + \ln \lambda_{km}\right)}{\sum_{m=1}^{M} \exp\left(\mathbf{x}_{obs,i} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km} - \frac{1}{2} \boldsymbol{\nu}'_{obs,i,km} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km} + \ln \lambda_{km}\right)} \\ &= \frac{\exp(\gamma_{km})}{\sum_{m=1}^{M} \exp(\gamma_{km})} \\ &= \omega_{imk} \end{aligned}$$

where

$$\gamma_{km} = \exp\left(\mathbf{x}_{obs,i} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km} - \frac{1}{2} \boldsymbol{\nu}'_{obs,i,km} \Xi_{k,obs,i}^{-1} \boldsymbol{\nu}'_{obs,i,km} + \ln \lambda_{km}\right), \quad (4.6)$$

and $\boldsymbol{\nu}_{obs,i,km}$ and $\Xi_{k,obs,i}$ are the mean and covariance matrix at level m for the continuous variables present for observation i. If the categorical variable, u_i, is observed, then the conditional probability ω_{imk} is just w_{im} . Thus,

$$E(z_{ik}w_{im} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}) = \begin{cases} \hat{z}_{ik}^{(t)}\omega_{imk} & \text{if } u_i \text{ is not observed,} \\ \hat{z}_{ik}^{(t)}w_{im} & \text{if } u_i \text{ is observed.} \end{cases}$$

Let

$$\hat{\omega}_{imk} = \begin{cases} \omega_{imk} & \text{if } u_i \text{ is not observed} \\ w_{im} & \text{if } u_i \text{ is observed} \end{cases}$$

for $m = 1, \ldots, M$ and $k = 1, \ldots, K$. Then we write

$$E(z_{ik}w_{im} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}) = \hat{z}_{ik}\hat{\omega}_{imk}$$

If v_{ij} is missing, let $\hat{v}_{ij;km} = E(v_{ij} \mid \mathbf{x}_{obs,i}\,,\, u_i = m; \theta_k^{(t)})$, the predicted value of v_{ij} from the regression in cell m of v_j on $\mathbf{v}_{obs,i}$, evaluated at the current value of the parameters, if observation i is in group k. If v_{ij} is observed, then $\hat{v}_{ij;km}$ is just v_{ij}. Thus we can write

$$\hat{v}_{ij;km} = \begin{cases} v_{ij} & \text{if } v_{ij} \text{ is present,} \\ E(v_{ij} \mid \mathbf{x}_{obs,i}\,,\, u_i = m; \theta_k^{(t)}) & \text{if } v_{ij} \text{ is missing.} \end{cases}$$

$$\begin{aligned} &E(z_{ik}w_{im}v_{ij} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}) \\ &\quad = E\left(z_{ik} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}\right) E\left(w_{im} \mid \mathbf{x}_{obs,i}\,; \boldsymbol{\phi}^{(t)}, z_{ik}\right) \\ &\qquad\qquad \times E\left(v_{ij} \mid \mathbf{x}_{obs,i}\,; w_{im}, z_{ik}, \boldsymbol{\phi}^{(t)}\right) \\ &\quad = \hat{z}_{ik}^{(t)}\hat{\omega}_{imk}\hat{v}_{ij;km} \end{aligned}$$

For $j,\ j' = 1, \ldots, p$,

$$\begin{aligned} &E\left(z_{ik}v_{ij}v_{ij'} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}\right) \\ &\quad = E\left(\sum_{m=1}^{M} w_{im}z_{ik}v_{ij}v_{ij'} \mid \mathbf{x}_{obs,i}\,; \boldsymbol{\phi}^{(t)}\right) \\ &\quad = E\left(z_{ik} \mid \mathbf{x}_{obs,i}; \boldsymbol{\phi}^{(t)}\right) \sum_{m=1}^{M} E\left(w_{im} \mid \mathbf{x}_{obs,i}\,; \boldsymbol{\phi}^{(t)}, z_{ik}\right) \\ &\qquad\qquad \times E\left(v_{ij}v_{ij'} \mid \mathbf{x}_{obs,i}\,; w_{im}, z_{ik}, \boldsymbol{\phi}^{(t)}\right) \end{aligned}$$

$$= \hat{z}_{ik}^{(t)} \sum_{m=1}^{M} \hat{\omega}_{imk} E\left(v_{ij}v_{ij'} \mid \mathbf{x}_{obs,i}\,; w_{im}, z_{ik}, \boldsymbol{\phi}^{(t)}\right)$$

$$= \begin{cases} \hat{z}_{ik}^{(t)} \sum_{m=1}^{M} \hat{\omega}_{imk} v_{ij} v_{ij'} & v_{ij},\ v_{ij'} \text{ both present} \\ \hat{z}_{ik}^{(t)} v_{ij'} \sum_{m=1}^{M} \hat{\omega}_{imk} \hat{v}_{ij;km} & v_{ij} \text{ missing, } v_{ij'} \text{ present} \\ \hat{z}_{ik}^{(t)} v_{ij} \sum_{m=1}^{M} \hat{\omega}_{imk} \hat{v}_{ij';km} & v_{ij} \text{ present, } v_{ij'} \text{ missing} \\ \hat{z}_{ik}^{(t)} \left\{ \sum_{m=1}^{M} \hat{v}_{ij;km} \hat{v}_{ij';km} + \sigma_{kjj';obs,i} \right\} & v_{ij},\ v_{ij'} \text{ both missing} \end{cases}$$

where $\sigma_{kjj';obs,i}$ denotes the conditional covariance of v_{ij} and v_{ij} given $\mathbf{x}_{obs,i}$, and that $u = m$, and the observation is in group k. (Note that it is an assumption of the location model that the covariance matrix is constant over the cells of the table.)

Note that these expressions are z_{ik} times the corresponding expressions for the general location model with data missing at random, described by Little and Schluchter (1985), and given by Little and Rubin (1987) on page 202.

The computations for the E step are easily performed by using the sweep operator introduced in section 4.3. Little and Schluchter (1985) give details on computations using the sweep operator for the location model with data missing at random. These are adapted as follows:

Suppose we let the matrix

$$Q = \begin{pmatrix} \hat{\Xi}_k & \hat{\Gamma}_k' \\ \hat{\Gamma}_k & P \end{pmatrix}$$

where P is a $M \times M$ diagonal matrix with the m^{th} element equal to $2 \ln \hat{\lambda}_{km}$ for $m = 1, \ldots, M$,

$$\hat{\Gamma}_k = \{\boldsymbol{\nu}_{km}\} = \begin{pmatrix} \nu_{k11} & \nu_{k21} & \nu_{kp1} \\ \nu_{k12} & \nu_{k22} & \nu_{kp2} \\ \vdots & \vdots & \vdots \\ \nu_{k1M} & \nu_{k2M} & \nu_{kpM} \end{pmatrix},$$

and $\hat{\Xi}_k$ is the variance covariance matrix for group k.

Suppose we partition $\hat{\Xi}_k$ and $\hat{\Gamma}_k$ such that

$$\hat{\Xi}_k = \begin{pmatrix} \hat{\Xi}_{k,obs,i} & \hat{\Xi}_{k,cov,i} \\ \hat{\Xi}_{k,cov,i} & \hat{\Xi}_{k,miss,i} \end{pmatrix}$$

and

$$\hat{\Gamma}_k = (\, \hat{\Gamma}_{k,obs,i}, \quad \hat{\Gamma}_{k,miss,i}\,),$$

where the current estimates of $\hat{\Xi}_k$ and $\hat{\Gamma}$ are partitioned according to the observed and missing variables in observation i. Sweeping on the observed elements of the

continuous variables observed for observation i yields the matrix

$$SWP[\mathbf{x}_{obs,i}]Q = \begin{pmatrix} G_{11} & G'_{12} & G'_{13} \\ G_{12} & G_{22} & G'_{23} \\ G_{13} & G_{23} & G_{33} \end{pmatrix}$$

where

$G_{11} = -\hat{\Xi}^{-1}_{k,obs,i}$

$G_{12} = \hat{\Xi}^{-1}_{k,obs,i}\hat{\Xi}_{k,cov,i}$, yields the regression coefficients of the missing x's on $\mathbf{x}_{obs,i}$,

$G_{13} = \hat{\Xi}^{-1}_{k,obs,i}\hat{\Gamma}_{k,obs,i}$, gives the coefficients of $\mathbf{x}_{obs,i}$ in the linear discriminant function (4.6),

$G_{22} = \hat{\Xi}_{k,miss,i} - \hat{\Xi}'_{k,cov,i}\hat{\Xi}^{-1}_{k,obs,i}\hat{\Xi}_{k,cov,i}$ contains the residual variances and covariances $\hat{\Xi}_{k,jj';obs,i}$ for x_{ij} and $x_{ij'} \in \mathbf{x}_{obs,i}$.

and the m^{th} diagonal element of $\frac{1}{2}G_{33} = \frac{1}{2}P - \frac{1}{2}\hat{\Gamma}_{k,obs,i}\hat{\Xi}^{-1}_{k,obs,i}\hat{\Gamma}'_{k,obs,i}$ yields the sum of the second and third terms of the expression in the brackets in (4.6).

In the M step of the algorithm, the new estimates $\theta^{(t+1)}$ of the parameters are estimated from the complete data sufficient statistics.

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ik}$$

$$\hat{\lambda}_{km} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{z}_{ik}\hat{\omega}_{imk}$$

$$\hat{\nu}_{kjm} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{z}_{ik}\hat{\omega}_{imk}\hat{v}_{ij;km}$$

$$\begin{aligned}\hat{\Xi}_{kjj'} &= \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{z}_{ik}E(w_{im}v_{ij}v_{ij'} \mid \mathbf{x}_{obs,i}, \theta^{(t)}_k) - \hat{\nu}_{kjm}\hat{\nu}_{kj'm} \\ &= \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{z}_{ik}\left[\sum_{m=1}^{M}\hat{\omega}_{imk}(\hat{v}_{ij;km} - \hat{\nu}_{kjm})(\hat{v}_{ij';km} - \hat{\nu}_{kj'm}) + C_{ki,jj'}\right]\end{aligned}$$

where, as in the mixture of multivariate normals described in section 4.5, an adjustment is required when both continuous variables,v_{ij} and $v_{ij'}$ are missing. The conditional covariance between variables j and j' for observation i, given that in group k, is defined as

$$C^{(t)}_{ki,jj'} = \begin{cases} 0 & \text{if either } v_{ij} \text{ or } v_{ij'} \text{ are observed,} \\ Cov(v_{ij}, v_{ij'} \mid \mathbf{x}_{obs,i}, \theta^{(t)}_k) & \text{if both } v_{ij} \text{ and } v_{ij'} \text{ are missing.} \end{cases}$$

The EM algorithm alternates between the E step and the M step until convergence.

4.8 Fitting multivariate mixture models with missing data

In section 4.5, we considered mixture models in which the component distributions were multivariate normal, and data were missing at random. In section 4.6, mixture models were considered for categorical variables with missing data values. In section 4.7, we considered missing data methods for mixture models in which the component distributions are location model distributions. In this section, we describe the general approach to multivariate mixture models for multivariate observations on both categorical and continuous variables where data are missing at random.

Suppose that p attributes are measured on n observations. We regard the data, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, as a random sample from a finite mixture of K component distributions

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

where $f_k(\mathbf{x}; \boldsymbol{\theta}_k)$ is the distribution for the kth component which has parameter vector $\boldsymbol{\theta}_k$, and $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$, for $k = 1, \ldots, K$.

Now, suppose that the vector of variables, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p)'$, has been partitioned as described in chapter 2, so that

$$\mathbf{x} = (\breve{\mathbf{x}}_1 | \ldots | \breve{\mathbf{x}}_l | \ldots | \breve{\mathbf{x}}_L)'$$

where we assume that within each of the K subpopulations, the variables within the partition cell $\breve{\mathbf{x}}_l$, are independent of the variables in cell $\breve{\mathbf{x}}_{l'}$, for $1 \leq l < l' \leq L$ (a form of local independence). The component distributions are thus of the form

$$f_k(\mathbf{x}; \boldsymbol{\theta}_k) = \prod_{l=1}^{L} f_{kl}(\breve{\mathbf{x}}_l; \boldsymbol{\theta}_{kl}).$$

where $\boldsymbol{\theta}_{kl}$ consists of the parameters of the distribution for partition cell l. Note that in each of the K classes, the random vector $\breve{\mathbf{x}}_l$ of partition cell l has the same type, however the parameters may vary from group to group.

The model for observation i, can thus be written as

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

As previously mentioned in Chapter 2, this thesis uses the following distributions for the partition cells $\breve{\mathbf{x}}_l$.

(a) *Discrete Distribution*

Where $\breve{\mathbf{x}}_l = \{x_j\}$ is a 1-dimensional discrete random variable taking values $1, \ldots, M_l$ with probabilities $\lambda_{kl1}, \ldots, \lambda_{klM_l}$.

(b) *Multivariate Normal*

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

(c) *Location Model*

Where $\breve{\mathbf{x}}_l$ is a $1+p_l$ dimensional vector of random variables with one discrete variable, u_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 $N_{p_l}(\boldsymbol{\nu}_{mkl}, \Xi_{kl})$.

To distinguish between the categorical and continuous variables, in this section, u_l will be used to denote the discrete variable when the lth partition cell is of type (a) or (c), and $\mathbf{v}_l$ will be used to denote the $p_l \times 1$ vector of continuous variables when the lth partition cell is of type (b) or (c). Thus, $\breve{\mathbf{x}}_l$ takes the form u_l, $\mathbf{v}_l$, and $(u_l, \mathbf{v}_l)$ when the lth partition cell is of type (a), (b) and (c) respectively.

Now suppose that data are missing at random throughout the dataset. The EM algorithm is applied to the mixture model, where the indicator vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are as described in section 2.2. There are two types of missing data that have to be considered in fitting the mixture model; one is the conceptual 'missing' data, the unobserved indicator of group membership, and the other is the unintended or accidental missing data values. However, these unintended missing values can be of four different types. They can be continuous and belong to a partition cell with either a multivariate normal component distribution or a location model component distribution, or they can be categorical variables involved in either a location model distribution or a discrete distribution.

The hypothetical 'complete data' consists of an $n \times p$ data array that includes the observed data and the values of the missing data, and the conceptual $n \times K$ array of class membership indicators.

The complete-data specification treats the $\mathbf{z}_i$ and the $\mathbf{x}_i$ as known leading to the loglikelihood

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

where

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

Maximising the loglikelihood $L_C(\phi)$ for the complete data is equivalent to maximising the loglikelihood $l_k(\boldsymbol{\theta}_k)$ separately for each partition cell, and maximising the $\sum\sum z_{ik}\log\pi_k$ with respect to the π_ks. The complete data sufficient statistics for the model are

(i) $\sum_{i=1}^{n} z_{ik}$, for each group k;

(ii) $\sum_{i=1}^{n} z_{ik}\delta_{ilm}$, for each class k, each categorical variable u_l, and each value m_l of u_l, where $\delta_{ilm} = \begin{cases} 1 & \text{if } x_{il} = m_l, \\ 0 & \text{otherwise.} \end{cases}$

(iii) Multivariate Normal partition cells

a. $\sum_{i=1}^{n} z_{ik}v_{ij}$ and $\sum_{i=1}^{n} z_{ik}v_{ij}^2$, for each group k, and each continuous variable v_j belonging to a multivariate normal partition cell.

b. $\sum_{i=1}^{n} z_{ik}v_{ij}v_{ij'}$, for each group k, and each pair of continuous variables, v_j and $v_{j'}$, $j < j'$, belonging to the same multivariate normal partition cell.

(iv) Location Model partition cells

a. $\sum_{i=1}^{n} z_{ik}w_{ilm}v_{ij}$ for each group k, each continuous variable v_j and each value m of the categorical variable u_l, belonging to the same location model partition cell;
where $w_{ilm} = \begin{cases} 1 & \text{if } u_l = m, \\ 0 & \text{otherwise.} \end{cases}$

b. $\sum_{i=1}^{n} z_{ik}v_{ij}v_{ij'}$, for each class k, each pair of continuous variables v_j and $v_{j'}$, $j \leq j'$, belonging to the same location model partition cell.

The E step of the EM algorithm requires the calculation of

$$Q(\phi, \phi^{(t)}) = E\{L_C(\phi) \mid \mathbf{x}_{obs}; \phi^{(t)}\},$$

the expectation of the complete data loglikelihood, conditional on the observed data and the current value of the parameters. We can calculate $Q(\phi, \phi^{(t)})$ by replacing z_{ik} with

$$\begin{aligned}\hat{z}_{ik} &= E(z_{ik} \mid \mathbf{x}_{obs,i}; \phi^{(t)}) \\ &= \frac{\pi_k f_k(\mathbf{x}_{obs,i}; \theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_{obs,i}; \theta_k^{(t)})}\end{aligned}$$

That is, z_{ik} is replaced by the estimate of the posterior probability $\hat{z}_{ik}$ that observation i belongs to group k. Depending on the model to be fitted, the E step also calculates for the appropriate partition cell,

(a) *Discrete Distribution*

- (*i*) $E(z_{ik}\delta_{ilm} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)})$.

  Refer to Section 4.6 for further details on this calculation.

(b) *Multivariate Normal*

- (*i*) $E(z_{ik}v_{ij} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)})$,
- (*ii*) $E(z_{ik}v_{ij}^2 \mid \mathbf{x}_{obs,i}, \theta_k^{(t)})$,
- (*iii*) $E(z_{ik}v_{ij}v_{ij'} \mid \mathbf{x}_{obs,i}; \theta_k^{(t)})$ for $j \neq j'$.

  Refer to Section 4.5 for further details on these calculations.

(c) *Location Model*

- (*i*) $E(z_{ik}w_{im} \mid \mathbf{x}_{obs,i}; \phi^{(t)})$
- (*ii*) $E(z_{ik}w_{im}v_{ij} \mid \mathbf{x}_{obs,i}; \phi^{(t)})$
- (*iii*) $E(z_{ik}v_{ij}v_{ij'} \mid \mathbf{x}_{obs,i}; \phi^{(t)})$.

  Refer to Section 4.7 for further details on these calculations.

At the M step of the algorithm, $\phi^{(t+1)}$ is chosen to be the value of ϕ that maximises $Q(\phi, \phi^{(t)})$ with respect to ϕ. For this model, the components of $\phi^{(t+1)}$ are given by

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ik}$$

(a) *Discrete Distribution estimates*

$$\hat{\lambda}_{klm} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n} \hat{z}_{ik}\hat{\delta}_{ilm}$$

for $k = 1, \ldots, K$, $l = 1, \ldots, L$ and $m = 1, \ldots, M_l$.

(b) *Multivariate Normal estimates*

$$\hat{\mu}_{kj}^{(t+1)} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \hat{v}_{ij,k}^{(t)}$$

$$\hat{\Sigma}_{kjj'}^{(t+1)} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[(\hat{v}_{ij,k}^{(t)} - \hat{\mu}_{kj}^{(t+1)})(\hat{v}_{ij',k}^{(t)} - \hat{\mu}_{kj'}^{(t+1)}) + C_{i,jj'}^{(t)} \right]$$

where

$$\hat{v}_{ij,k}^{(t)} = \begin{cases} v_{ij} & \text{if } v_{ij} \text{ is present} \\ E(v_{ij} \mid \mathbf{v}_{obs,i}, \theta_k^{(t)}) & \text{if } v_{ij} \text{ is missing,} \end{cases}$$

and where the residual covariance

$$C_{ki,jj'}^{(t)} = \begin{cases} 0 & \text{if either } v_{ij} \text{ or } v_{ij'} \text{ are observed} \\ Cov(v_{ij}, v_{ij'} \mid \mathbf{v}_{obs,i}, \theta_k^{(t)}) & \text{if both } v_{ij} \text{ and } v_{ij'} \text{ are missing.} \end{cases}$$

for $k = 1, \ldots, K$, $l = 1, \ldots, L$ and $j, \quad j' \in l$.

(c) *Location Model estimates*

$$\lambda_{klm} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \hat{\omega}_{imk}^{(t)}$$

$$\hat{\nu}_{kjm} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik} \hat{\omega}_{imk}^{(t)} \hat{v}_{ij;km}$$

$$\begin{aligned} \hat{\Xi}_{kjj'} &= \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik} E(w_{im} v_{ij} v_{ij'} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) - \hat{\nu}_{kjm} \hat{\nu}_{kj'm} \\ &= \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik} \left[\sum_{m=1}^{M} \hat{\omega}_{imk} (\hat{v}_{ij;km} - \hat{\nu}_{kjm})(\hat{v}_{ij';km} - \hat{\nu}_{kj'm}) \right. \\ &\qquad \left. + C_{ki,jj'} \right] \end{aligned}$$

where

$$C_{ki,jj'}^{(t)} = \begin{cases} 0 & \text{if either } v_{ij} \text{ or } v_{ij'} \text{ are observed,} \\ Cov(v_{ij}, v_{ij'} \mid \mathbf{x}_{obs,i}, \theta_k^{(t)}) & \text{if both } v_{ij} \text{ and } v_{ij'} \text{ are missing,} \end{cases}$$

and

$$\hat{v}_{ij;km} = \begin{cases} v_{ij} & \text{if } v_{ij} \text{ is present,} \\ E(v_{ij} \mid \mathbf{x}_{obs,i}, u_i = m; \theta_k^{(t)}) & \text{if } v_{ij} \text{ is missing.} \end{cases}$$

for $k = 1, \ldots, K$, $l = 1, \ldots, L$, and $j, \quad j' \in l$.

The EM algorithm alternates between the E step and the M step until convergence.

4.9 Implementing MULTIMIX with missing data

The program requires the data to be in the form of an $n \times p$ matrix of observations by variables, where missing data has been coded to the value currently set in the Fortran parameter statement. This parameter, *imiss*, is currently set at -999.

The iterations may be started either from an initial classification or from an initial set of parameter estimates. A program has been written to create the parameter input file needed for program MULTIMIX. Details of the parameter input file for the program can be found in Appendix 1.

The current version of the program written for this thesis to implement the approach of multivariate finite mixture models where data are missing at random, uses a convergence criterion to cease iterating when the difference in loglikelihoods at iteration t and iteration $(t - 10)$ is 1.0×10^{-10}.

Little and Schlucter (1985) in their paper on the maximum likelihood estimation of the parameters of the distribution for a location model where data are missing at random, suggest grouping together observations with common missing data patterns, to avoid unnecessary sweep operations and save computational time. The program written for this thesis, does not follow this suggestion because, as each observation vector is partitioned into L partition cells, the ordering required to find common missing data patterns within each partition cell followed by the subsequent reordering, would increase the computational time (and the memory requirements of the program), over that of performing some unnecessary sweep operations. Other possibilities were considered to avoid unnecessary sweep operations, but these would increase the memory requirements of the program. Also, with highly multivariate data, there will be a large number of missing data patterns with small numbers of observations with each type.

For each observation, the partition cells are examined for the presence of missing data. In the multivariate normal and location model partition cells, the matrices for sweeping are created using the current parameter estimates in the manner described in sections 4.5 and 4.7 respectively, if data are missing. These matrices are swept on the observed variables for that partition cell and the conditional means and covariances are calculated for imputing at the M step. The estimates of the posterior probabilities are calculated for each observation. The program then proceeds to the M step, where the parameters are calculated as given in section 4.8.

In the next chapter, we illustrate the approach of multivariate finite mixture models by clustering multivariate data sets where data are missing at random.

Chapter 5

Clustering with missing data

5.1 Introduction

In this chapter, we demonstrate that mixture model likelihood methods can be used to usefully cluster multivariate data sets where the variables can be either categorical or continuous, and where some of the data values are missing at random in the sense of Little & Rubin (1987). In section 5.2, we illustrate the approach of multivariate finite mixture models for a data set in which the underlying structure is known. The extent of recovery of the true structure known to exist in this data set is initially evaluated for two well separated clusters, and then for two clusters that are more difficult to distinguish between. In section 5.3, we illustrate the scope of the methods proposed in Chapter 4 for the pretrial variables of the full cancer data set.

5.2 The data set

The iris data previously clustered in section 3.3 was chosen to illustrate the approach of multivariate finite mixture models with missing data for a multivariate normal distribution. As all four variables are observed for all observations, missing data were created, with the probability of an observation on a variable being missing taken as $1/3$ independently of all other data values. The missing values generated in this fashion are missing completely at random, and the missing data mechanism is ignorable for likelihood based inferences (Refer to section 4.1). As the amount of missing data is fairly extreme, it should be a good test of the models ability to recover the distributional structure in this classical data set.

In the iris data set, it is known that the species *I. Versicolor* and *I. Virginica* are more similar to each other than either is to the *I. Setosa* (see section 3.3 for further details and references). Initially we shall consider the clustering of cases of the two well separated clusters, *I. Setosa* and *I. Versicolor*, where missing data values have been created as per above. We shall then look at the clustering of the two more similar species, *I. Versicolor* and *I. Virginica*, with missing data values. The clusters found under the various models fitted will be compared with the species classification.

Similar notation to that previously defined in section 3.2.1 will be used. To distinguish between models previously used for the iris data set, and the models to

be used for the two, well separated species of the iris data with missing data, 'Separate,' will be included in the description of all models for the *I. Setosa* and *I. Versicolor* data set with missing data values. The models to be used for the two, more similar species of the iris data, *I. Versicolor* and *I. Virginica* with missing data values, will have 'Close, ' included in the description of all models. As only models with two groups will be investigated, the number of groups to be fitted to the models will be omitted from the notation. Hence, 'Separate, Model i' denotes the model that has partitioning i and two groups, fitted for the two species of iris, *I. Setosa* and *I. Versicolor*, with missing data values. Similarly, 'Close, Model i' denotes the model that has partitioning i and two groups, fitted for *I. Versicolor* and *I. Virginica* with missing data values.

5.2.1 Well separated clusters

The data set to be clustered is displayed in Table 5.1, and consists of four variables measured on each of 50 observations from the species *I. Setosa* and *I. Versicolor*. There are 144 missing values scattered at random throughout the data set. It can be seen that one observation has all four variables missing, thirteen observations have three variables recorded as missing, 35 observations have two variables recorded as missing, 31 observations have one variable recorded as missing, and 20 observations have no variables recorded as missing.

The available values for each variable were used to examine the distributions for each variable. The distributions appeared approximately normally distributed within the mixture population, however petal length and petal width were both bi modal distributions.

Table 5.1

Data for Separated, Models 1 and 2.

	I. Setosa										*I. Versicolor*									
Obs	x_1	x_2	x_3	x_4	**Obs**	x_1	x_2	x_3	x_4	**Obs**	x_1	x_2	x_3	x_4	**Obs**	x_1	x_2	x_3	x_4
1	5.1	3.5	*	*	**26**	5.0	3.0	*	0.2	**51**	7.0	*	*	1.4	**76**	*	3.0	4.4	1.4
2	4.9	*	1.4	0.2	**27**	*	3.4	1.6	0.4	**52**	6.4	3.2	4.5	1.5	**77**	6.8	*	4.8	*
3	*	3.2	1.3	*	**28**	5.2	3.5	*	0.2	**53**	*	3.1	4.9	1.5	**78**	*	3.0	*	1.7
4	4.6	*	*	0.2	**29**	5.2	3.4	*	0.2	**54**	*	2.3	4.0	*	**79**	6.0	2.9	*	*
5	*	*	1.4	*	**30**	4.7	3.2	1.6	0.2	**55**	6.5	*	4.6	*	**80**	*	2.6	*	1.0
6	5.4	*	1.7	0.4	**31**	*	3.1	1.6	0.2	**56**	5.7	2.8	4.5	*	**81**	*	*	3.8	*
7	4.6	3.4	1.4	0.3	**32**	5.4	3.4	1.5	0.4	**57**	*	*	4.7	1.6	**82**	5.5	2.4	3.7	*
8	*	3.4	*	0.2	**33**	5.2	4.1	*	*	**58**	*	2.4	3.3	*	**83**	5.8	*	3.9	*
9	4.4	*	1.4	*	**34**	5.5	4.2	1.4	0.2	**59**	*	*	*	1.3	**84**	6.0	2.7	5.1	1.6
10	*	*	1.5	0.1	**35**	4.9	*	*	*	**60**	5.2	*	*	*	**85**	5.4	*	4.5	1.5
11	5.4	*	1.5	0.2	**36**	*	3.2	*	*	**61**	*	2.0	3.5	1.0	**86**	6.0	*	4.5	1.6
12	4.8	*	*	0.2	**37**	5.5	*	*	*	**62**	*	3.0	4.2	*	**87**	6.7	*	4.7	1.5
13	4.8	3.0	1.4	0.1	**38**	4.9	3.6	1.4	0.1	**63**	6.0	*	*	*	**88**	6.3	2.3	4.4	1.3
14	*	*	*	0.1	**39**	4.4	3.0	1.3	0.2	**64**	6.1	2.9	4.7	1.4	**89**	5.6	3.0	4.1	*
15	*	*	1.2	0.2	**40**	5.1	*	*	0.2	**65**	5.6	*	*	1.3	**90**	5.5	*	4.0	*
16	*	*	*	*	**41**	5.0	3.5	*	0.3	**66**	*	3.1	4.4	1.4	**91**	*	*	*	1.2
17	5.4	3.9	1.3	0.4	**42**	4.5	2.3	1.3	0.3	**67**	5.6	*	*	1.5	**92**	6.1	3.0	4.6	1.4
18	*	3.5	1.4	0.3	**43**	*	3.2	1.3	0.2	**68**	*	2.7	*	1.0	**93**	5.8	*	4.0	*
19	5.7	3.8	1.7	0.3	**44**	5.0	*	1.6	0.6	**69**	6.2	2.2	*	1.5	**94**	5.0	2.3	3.3	1.0
20	*	3.8	1.5	*	**45**	5.1	3.8	*	0.4	**70**	5.6	*	*	*	**95**	5.6	2.7	4.2	1.3
21	5.4	*	1.7	*	**46**	4.8	*	*	0.3	**71**	5.9	3.2	4.8	*	**96**	*	3.0	4.2	1.2
22	5.1	3.7	*	0.4	**47**	5.1	*	*	*	**72**	6.1	2.8	*	*	**97**	5.7	2.9	*	*
23	4.6	3.6	1.0	*	**48**	4.6	3.2	1.4	0.2	**73**	6.3	2.5	4.9	1.5	**98**	*	2.9	*	1.3
24	*	*	*	0.5	**49**	5.3	*	1.5	0.2	**74**	*	2.8	4.7	1.2	**99**	5.1	2.5	*	*
25	*	3.4	1.9	*	**50**	5.0	3.3	1.4	0.2	**75**	6.4	2.9	4.3	*	**100**	*	*	4.1	1.3

* denotes missing value

x_1 is Sepal length, x_2 is Sepal width, x_3 is Petal length, and x_4 is Petal width.

The model of complete local independence will be fitted initially. The component distributions will be the $N(\mu_{kl}, \sigma^2_{kl})$ density function for each of the 4 continuous variables for this model. The model with complete local independence fitted to this data set will be referred to as Separate, Model 1.

The model was fitted iteratively using the EM algorithm from initial estimates of the parameters based on the available cases of the observations with a petal length was less than 2, and at least 2. In order to search for other maxima, the model was also fitted from other classifications generated by splitting the observations into two groups using various criteria. From the starting values tried, all converged to a solution where the loglikelihood was -94.0092, the same solution that was found using initial parameter estimates based on the species classification. The observations were assigned to their group of greater probability for Separate, Model 1.

Table 5.2

Agreements and differences between the species and the model classifications for Separate, Model 1.

Species	Classification Model	
	Group 1	**Group 2**
I. Setosa	48	2
I. Versicolor	1	49

It can be seen from Table 5.2 that there are three observations where the species and the model classifications differ. An examination of the posterior probabilities showed that 91 of the 100 observations are definitely assigned to a group. The three observations that the model classifies in a different group from the one corresponding to the species, have posterior probabilities of assignment to Group 1 (equivalent to *I. Setosa*), of $\hat{\tau}_{16,1} = 0.4989$, $\hat{\tau}_{37,1} = 0.4344$, and $\hat{\tau}_{60,1} = 0.8106$. It can be seen that observations 16 and 37, both *I. Setosa* plants, have not really been decisively assigned to either group. It can be seen from Table 5.1 that observation 16 has four missing values, whilst observations 37 and 60 both have only Sepal length present. A comparison of the Sepal length with the parameter estimates in Table 5.3 indicates that observation 60 has a sepal length that is more similar to that of *I Setosa* species.

The parameter estimates for Separate, Model 1 are displayed in Table 5.3. The estimates of the mixing proportions in Groups 1 and 2 are 0.4989 and 0.5011,

leading to expected numbers in the two groups respectively of 49.89 and 50.11. This is very close to the 50/50 division in the species.

Table 5.3

Parameter estimates for Separate, Model 1.

Variable	Group	Mean	STDEV
Sepal length	1	5.021	0.332
	2	5.939	0.464
Sepal width	1	3.412	0.379
	2	2.752	0.315
Petal length	1	1.453	0.173
	2	4.312	0.449
Petal width	1	0.258	0.114
	2	1.359	0.190

The parameter estimates displayed in Table 5.3 may be compared with those according to the species classification displayed in Table 3.19. It can be seen that the estimates are fairly close.

The group assignment for Separate, Model 1 was used and the within group correlation structure was examined. As is to be expected, significant correlations ($r^* = 0.278$ for a sample size of 50 at the 0.05 level of significance), existed between each pair of variables in at least one of the groups. All four variables were grouped together into one partition, allowing for a general correlation pattern.

A mixture of two component distributions will be fitted, where each component is the $N_4(\boldsymbol{\mu}_{k1}, \Sigma_{k1})$ density. We will refer to this model as Separate, Model 2.

The model was fitted using initial parameter estimates based on the available cases of the observations where the petal length was at least 2, and less than 2. In order to search for other maxima, the model was fitted from other classifications generated by splitting the observations into two groups using other criteria. The starting values tried converged to a solution where the loglikelihood was -58.382, the same solution found when using the group assignment based on the species classification or the Separate, Model 1 assignments.

When the observations were assigned to their group of greater probability for Separate, Model 2, exactly the same classification results as that given for Separate, Model 1, and displayed in Table 5.2. An examination of the posterior probabilities showed that 93 of the observations are definitely assigned to a group. The three observations that this model classifies into a different group from the one

corresponding to the species, are the same three observations that were also misclassified under Separate, Model 1. These observations are assigned to Group 1 with posterior probabilities of $\hat{\tau}_{16,1} = 0.4936$, $\hat{\tau}_{37,1} = 0.3985$, and $\hat{\tau}_{60,1} = 0.7728$.

The parameter estimates for Separate, Model 2 are displayed in Table 5.4. The estimates of the mixing proportions in each group for this model are 0.4936 and 0.5064.

Table 5.4

Parameter estimates for Separate, Model 2.

Mean vectors:

$$\hat{\boldsymbol{\mu}}_{11} = (5.012,\ 3.420,\ 1.455,\ 0.2601)$$

$$\hat{\boldsymbol{\mu}}_{21} = (5.895,\ 2.761,\ 4.291,\ 1.326)$$

Variance covariance matrices:

$$\hat{\Sigma}_{11} = \begin{pmatrix} 0.107 & & & \\ 0.074 & 0.133 & & \\ 0.027 & 0.006 & 0.027 & \\ 0.012 & 0.007 & 0.006 & 0.013 \end{pmatrix},$$

and

$$\hat{\Sigma}_{21} = \begin{pmatrix} 0.238 & & & \\ 0.061 & 0.095 & & \\ 0.156 & 0.066 & 0.206 & \\ 0.045 & 0.028 & 0.073 & 0.036 \end{pmatrix}.$$

where the variables are
Sepal length, Sepal width, Petal length, and Petal width.

The improvement in fit gained by the addition of the covariances to Separate, Model 1 was investigated. Separate, Model 2 requires an additional twelve parameters to be estimated in comparison to those required to be estimated for Separate, Model 1. Twice the difference in the loglikelihoods is 71.254. Using the log-likelihood ratio as a guide, it can be seen that there is a definite improvement in using Separate, Model 2 in preference to Separate, Model 1. There is also a slight increase in the numbers of observations definitely assigned to a group. However convergence was slightly slower than for Separate, Model 1.

Table 5.5

Parameter Estimates for the Iris Data using the Species Classifications

Mean Vectors:

$$\hat{\boldsymbol{\mu}}_{11} = (5.006,\ 3.428,\ 1.462,\ 0.246)$$

$$\hat{\boldsymbol{\mu}}_{21} = (5.936,\ 2.770,\ 4.260,\ 1.326)$$

$$\hat{\boldsymbol{\mu}}_{31} = (6.588,\ 2.974,\ 5.552,\ 2.026)$$

Variance Covariance Matrices:*

$$\hat{\Sigma}_{11} = \begin{pmatrix} 0.122 & & & \\ 0.097 & 0.141 & & \\ 0.016 & 0.011 & 0.030 & \\ 0.010 & 0.009 & 0.006 & 0.011 \end{pmatrix}$$

$$\hat{\Sigma}_{21} = \begin{pmatrix} 0.261 & & & \\ 0.083 & 0.096 & & \\ 0.179 & 0.081 & 0.216 & \\ 0.055 & 0.040 & 0.072 & 0.038 \end{pmatrix}$$

$$\hat{\Sigma}_{31} = \begin{pmatrix} 0.396 & & & \\ 0.092 & 0.102 & & \\ 0.297 & 0.070 & 0.298 & \\ 0.048 & 0.047 & 0.048 & 0.074 \end{pmatrix}$$

where the variables are
Sepal length, Sepal width, Petal length, and Petal width,
Group 1 is *I Setosa*, Group 2 is *I Versicolor*, and Group 3 is *I Virginica*.

Another comparison between the species classification and the Separate Model 2 fit can be obtained by comparing the estimated parameters for Separate Model 2 with their counterparts using the species classification. It can be seen from a comparison of Tables 5.4 and 5.5 that the agreement is fairly close.

The estimates of the variance covariance matrices for the model appear to be reasonable, and as is expected, they were always positive definite. The iris data set was taken and variables were made missing, where the probability of a variable being missing was 1/3. MINITAB was used to calculate estimates of the variance covariance matrices and the eigenvalues of these estimates were then calculated. This procedure was repeated 300 times. MINITAB calculated estimates of the variance covariance matrices for *I. Virginica* in which 28.7% of the estimates had

* Calculated using divisor of n.

eigenvalues of less than zero. For *I. Setosa*, 15% of the estimates also had eigenvalues less than zero. MINITAB calculates the covariances between each pair of variables (refer to MINITAB reference manual). If some data are missing, the covariance between each pair of variables is calculated using all rows that have both values present. This can lead to variance covariance matrices that are not positive definite. If the variance covariances were calculated using only those observations for which all four variables were observed, the covariances would be calculated using a very small amount of data, in this example on twelve *I. Setosa* observations and eight *I. Versicolor* observations.

As has been previously pointed out, equal covariance matrices cannot be imposed for the component distributions in the current version of the program. It may be an advantage to impose this condition on this data set, as this would reduce the numbers of parameters to be estimated from the small data set.

It has been demonstrated that with 36% of the data values missing, the approach of multivariate finite mixture models with data missing at random has been able to detect the structure present in the data exceedingly well when the clusters are well separated. In the next section, we will investigate the performance of the model when it is known that the data consists of two clusters that are closer together.

Table 5.6

Data for Close, Models 1 and 2.

	I. Versicolor									*I. Virginica*								
Obs	x_1	x_2	x_3	x_4	**Obs**	x_1	x_2	x_3	x_4	**Obs**	x_1	x_2	x_3	x_4	**Obs**	x_1	x_2	x_3	x_4
1	7.0	*	4.7	*	**26**	6.6	*	*	1.4	**51**	6.3	*	6.0	*	**76**	7.2	3.2	*	*
2	6.4	3.2	*	1.5	**27**	*	*	4.8	*	**52**	5.8	2.7	5.1	1.9	**77**	*	2.8	4.8	1.8
3	*	*	*	1.5	**28**	6.7	3.0	5.0	1.7	**53**	7.1	3.0	5.9	*	**78**	*	3.0	4.9	1.8
4	5.5	2.3	4.0	*	**29**	*	2.9	4.5	*	**54**	*	2.9	5.6	1.8	**79**	6.4	*	5.6	2.1
5	6.5	*	*	*	**30**	5.7	2.6	3.5	*	**55**	6.5	*	*	*	**80**	*	3.0	5.8	1.6
6	5.7	*	4.5	1.3	**31**	*	2.4	*	*	**56**	7.6	3.0	6.6	*	**81**	*	2.8	6.1	1.9
7	6.3	3.3	*	1.6	**32**	5.5	2.4	*	1.0	**57**	4.9	2.5	4.5	1.7	**82**	7.9	3.8	*	*
8	4.9	2.4	3.3	1.0	**33**	5.8	2.7	*	*	**58**	*	2.9	6.3	1.8	**83**	*	*	5.6	2.2
9	6.6	2.9	*	*	**34**	6.0	*	*	1.6	**59**	*	*	5.8	*	**84**	*	*	*	1.5
10	5.2	2.7	3.9	1.4	**35**	5.4	3.0	4.5	1.5	**60**	7.2	3.6	6.1	2.5	**85**	*	*	*	1.4
11	5.0	2.0	3.5	1.0	**36**	6.0	3.4	*	1.6	**61**	*	3.2	5.1	*	**86**	7.7	*	6.1	2.3
12	*	3.0	*	*	**37**	*	3.1	4.7	*	**62**	*	*	5.3	1.9	**87**	6.3	*	5.6	2.4
13	*	2.2	*	*	**38**	6.3	2.3	4.4	*	**63**	6.8	*	5.5	2.1	**88**	6.4	3.1	*	*
14	6.1	2.9	*	1.4	**39**	*	3.0	*	*	**64**	*	*	5.0	*	**89**	*	3.0	4.8	*
15	5.6	2.9	*	*	**40**	*	*	4.0	1.3	**65**	*	*	5.1	2.4	**90**	6.9	*	*	2.1
16	*	3.1	4.4	1.4	**41**	5.5	*	4.4	1.2	**66**	6.4	3.2	5.3	*	**91**	6.7	3.1	5.6	2.4
17	*	3.0	4.5	*	**42**	6.1	3.0	4.6	1.4	**67**	*	*	*	*	**92**	*	3.1	*	2.3
18	*	*	4.1	*	**43**	5.8	2.6	*	*	**68**	7.7	3.8	6.7	2.2	**93**	5.8	2.7	5.1	*
19	*	2.2	4.5	1.5	**44**	5.0	2.3	3.3	1.0	**69**	7.7	2.6	*	2.3	**94**	*	3.2	5.9	*
20	*	*	3.9	*	**45**	*	*	4.2	1.3	**70**	6.0	2.2	5.0	1.5	**95**	6.7	*	5.7	2.5
21	5.9	*	*	1.8	**46**	*	*	*	1.2	**71**	*	3.2	5.7	*	**96**	*	3.0	*	2.3
22	6.1	2.8	4.0	1.3	**47**	*	2.9	*	1.3	**72**	*	2.8	*	2.0	**97**	*	2.5	5.0	1.9
23	6.3	2.5	*	*	**48**	6.2	2.9	4.3	1.3	**73**	*	2.8	6.7	*	**98**	6.5	3.0	5.2	2.0
24	*	2.8	*	1.2	**49**	5.1	2.5	3.0	1.1	**74**	*	2.7	4.9	1.8	**99**	6.2	3.4	5.4	*
25	6.4	2.9	*	*	**50**	*	*	4.1	*	**75**	*	3.3	5.7	2.1	**100**	5.9	3.0	*	*

* denotes missing value

x_1 is Sepal length, x_2 is Sepal width, x_3 is Petal length, and x_4 is Petal width.

5.2.2 Poorly separated clusters

The data set to be clustered is displayed in Table 5.6. It consists of the four variables measured on each of 50 observations from the species *I. Versicolor* and *I. Virginica*. There are 152 missing data values scattered at random throughout the data set. It can be seen from Table 5.6 that there is one observation that has all four variables missing, sixteen observations have three variables recorded as missing, 34 observations have two variables recorded as missing, 32 observations have one variable recorded as missing and seventeen observations have no variables recorded as missing.

The available values for each variable were used to examine the distributions for each variable. The distributions appeared approximately normal within the mixture population.

The model of complete independence will be fitted initially. The model of complete independence fitted to this data set will be referred to as Close, Model 1. The model was fitted iteratively with the EM algorithm, where the initial estimates of the parameters were based on the available cases of the observations with a petal width of no more than 1.6, and at least 1.7. The model was also fitted from classifications generated by splitting the observations into two groups using various criteria. From the starting values tried, all converged to a solution where the loglikelihood was -161.035. This is the same solution that was found when using the initial grouping based on the species classification.

The observations were assigned to their group of greater probability for Close, Model 1.

Table 5.7

Agreements and differences between the species and the model classifications for Close, Model 1.

	Classification	
Species	**Model**	
	Group 1	**Group 2**
I. Versicolor	37	13
I. Virginica	6	44

It can be seen from Table 5.7 that there are nineteen observations where the species and the model classifications differ. An examination of the posterior probabilities showed that 64 of the 100 observations were definitely assigned to a group, and 19 observations had posterior probabilities of assignment to a group lying between 0.5

and 0.8. The posterior probabilities of the observations that are misclassified by the model were examined. Three observations were definitely assigned to a group that is different to the one that corresponds to the species. These three observations had all four variables recorded for each observation. Three observations were not decisively assigned to either group for Close, Model 1. It must be remembered that in the analogous model that was fitted to the iris data set with no missing data values, that is, Iris Model 1, there were 9 observations misclassified in the Versicolor and the Virginica species.

The parameter estimates for Close, Model 1 are displayed in Table 5.8. The estimates of the mixing proportions in each group are 0.432 and 0.568, leading to expected numbers in each group of 43.2 and 56.8.

Table 5.8

Parameter estimates for Close, Model 1.

Variable	Group	Mean	STDEV
Sepal length	1	5.703	0.485
	2	6.691	0.579
Sepal width	1	2.643	0.317
	2	3.047	0.289
Petal length	1	4.178	0.516
	2	5.520	0.563
Petal width	1	1.329	0.199
	2	2.002	0.300

The parameter estimates displayed in Table 5.8 were compared with their counterparts for the species classification, displayed in Table 3.19. It can be seen that the estimates are fairly close.

It is known that significant correlations exist between each pairs of variables in at least one of the groups. All four variables were then grouped together into one partition allowing for a general correlation pattern.

A mixture of two component distributions will be fitted, where each component is the $N_4(\boldsymbol{\mu}_{k1}, \Sigma_{k1})$ density. We will refer to this model as Close, Model 2.

The model was fitted iteratively using the EM algorithm with an initial grouping based on the classification assignments from Close, Model 1. In order to search for other maxima, the model was also fitted from other classifications generated by splitting the observations into two groups using other criteria. Two solutions of the likelihood equation were found. The solution corresponding to the largest of the local maxima was used, the solution where the loglikelihood was -114.736. This

is the same solution that was found using the species classifications as the initial grouping. The observations were assigned to their group of greater probability for Close, Model 2.

Table 5.9

Agreements and differences between the species and the model classifications for Close, Model 2.

Species	Classification Model	
	Group 1	**Group 2**
I. Versicolor	36	14
I. Virginica	6	44

It can be seen from Table 5.9 that there are twenty differences between the model and the species classifications. An examination of the posterior probabilities showed that 62 of the 100 observations were definitely assigned to a group. The observations that were misclassified by the model were examined. Four of these observations were definitely assigned to a different group than the one corresponding to the species, and seven observations were not decisively assigned to either Group 1 or Group 2.

An investigation of the differences in the cluster assignments between Close Models 1 and 2, showed that there had been quite a change in the cluster assignments, with seventeen observations being assigned to a different group. Four observations were definitely assigned to a different group by both models, and six observations were definitely assigned to a different group by either model. Sixteen of these observations had at least one variable missing.

The parameter estimates for Close, Model 2 are displayed in Table 5.10. The estimates of the mixing proportions are 0.4216 and 0.5784, leading to expected numbers in each group of 42.2 and 57.8.

Table 5.10

Parameter estimates for Close, Model 2.

Mean vectors:

$$\hat{\boldsymbol{\mu}}_{11} = (5.879,\ 2.858,\ 4.270,\ 1.324)$$

$$\hat{\boldsymbol{\mu}}_{21} = (6.529,\ 2.911,\ 5.440,\ 1.990)$$

Variance covariance matrices:

$$\hat{\Sigma}_{11} = \begin{pmatrix} 0.250 & & & \\ 0.117 & 0.098 & & \\ 0.192 & 0.143 & 0.309 & \\ 0.054 & 0.050 & 0.083 & 0.031 \end{pmatrix},$$

and

$$\hat{\Sigma}_{21} = \begin{pmatrix} 0.457 & & & \\ 0.132 & 0.138 & & \\ 0.350 & 0.142 & 0.410 & \\ 0.110 & 0.075 & 0.119 & 0.098 \end{pmatrix}.$$

where the variables are
Sepal length, Sepal width, Petal length, and Petal width,
Group 1 corresponds to *I Versicolor*, Group 2 corresponds to *I Virginica*.

Another comparison between the species classification and the Close Model 2 fit can be obtained by comparing the estimated parameters for Close Model 2 with their counterparts using the species classification. It can be seen from a comparison of Tables 5.10 and 5.5 that the agreement is fairly close.

The improvement in fit gained by the addition of the covariances to Model 1 was investigated. It has been pointed out in section 5.2.1 that the partitioning for Model 2 requires an additional twelve parameters to be estimated in comparison to those required to be estimated for Model 1. For the data set analysed in this section, twice the difference in the loglikelihoods is 92.598. Using the log-likelihood ratio as a guide, there is clearly an improvement in using Close, Model 2 in preference to Close, Model 1. However, for Close, Model 2, there has been an increase in the numbers of observations that are not decisively assigned to either group. (This model has 9 observations that are not decisively assigned to either group whilst Close, Model 1 has 3 in this category). There has also been a slight decrease in the numbers of observations that are definitely assigned to a group. Convergence was slower for Close, Model 2 than for Close, Model 1.

It has been demonstrated in section 5.2.1 that the approach of multivariate finite mixture models has been able to detect the structure known to exist in a data set when the amount of missing data is fairly extreme, and the two clusters are well separated. In section 5.2.2, the approach has been applied to the data set containing only the species, *I. Versicolor* and *I. Virginica*, species in which there is overlap, again with the data set having a fairly extreme amount of missing data values created. These two clusters are more difficult to distinguish between even when there are no data values missing than are the *I Setosa* and *I. Versicolor* clusters analysed in section 5.2.1. The approach of multivariate finite mixture models

has been able to detect the known structure in these two poorly separated clusters, and has been able to compute reasonable estimates of the variance covariance matrices.

In the next section, the scope of the methods that have been proposed in Chapter 4 will be illustrated by considering the clustering of cases on the basis of the pretrial variables for the cancer data set. In chapter 3, the cancer data set was clustered using only those observations that had no missing values in any of the pretrial covariates. In the following section, we shall cluster the full data set which shall now include those observations that were omitted from the previous analyses.

5.3 The cancer data

The full data set consists of 506 observations, where 31 observations have one or more missing data values in the twelve pretrial variables. In total, there are 62 missing data values scattered throughout the data set. Some of these missing data values are categorical variables, and some are continuous. We will assume that these missing data values are missing at random. Further details on the cancer data set can be found in section 3.2.

Models fitted to the full cancer data set will have 'Full' included in the description of all models. This will enable us to distinguish between analogous models fitted for the full cancer data set and the data set previously analysed in section 3.2. 'Full Model i' denotes the model that has partitioning i and 2 groups, and is fitted to the full cancer data set.

The general strategy followed in Chapter 2 and described in section 3.4.2 for fitting the model to the data will not be observed in this instance, because the additional information from the 31 incomplete observations should not greatly affect the clustering. We will therefore fit a two group model with the same partitioning as was used for Model 3, that is the model in which the variables Wt, SBP and DBP are grouped together in a partition. This model will be referred to as Full Model 3.

The model was fitted iteratively using the EM algorithm from initial estimates of the parameters based on those resulting from the clinical classification, and also from initial classifications generated by randomly splitting the observations into two groups. Three solutions of the likelihood equation were found. From ten starting values, seven converged to a solution where the loglikelihood was -11895.758, the solution that was found using the parameter estimates based on the clinical classification. Two starting values converged to a solution where the loglikelihood was -11980.677 and one converged to a solution where the loglikelihood was

-11902.775. The solution corresponding to the largest of the local maxima was used.

Table 5.11

Agreements and differences between the Clinical and Full Model 3 classifications.

		Model Classification	
	Class	**Group 1**	**Group 2**
Clinical Classification	**Stage 3**	270	22
	Stage 4	21	193

When the observations are assigned to their group of greater probability for Full Model 3, it can be seen from Table 5.11 that the clinical classification and the 'statistical diagnosis' are different for 43 observations. An examination of the posterior probabilities showed that 421 of the 506 observations are definitely assigned† to a group.

It can be seen from a comparison of Table 3.5 and Table 5.11, that the numbers of observations where the clinical and the model classifications differ, are very similar. An investigation of the differences in the group assignments for the 475 observations that were clustered for Model 3 showed that three of these observations were classified differently for Full Model 3. The three observations, observations 58, 220 and 498 are all assigned for Full Model 3, to Group 1 (equivalent to the clinical classification of Stage 3) with posterior probabilities $\hat{\tau}_{i1}$ of 0.563, 0.501 and 0.560 respectively. As can be seen, none of these three observations is decisively assigned to a group. For Model 3, observations 58 and 220 were also not decisively assigned to a group (see Table 3.7).

The logarithm of the odds, $p/(1-p)$ where in this instance, p is the posterior probability of being assigned to Group 1, was calculated for the 475 observations that were clustered for both Model 3 and Full Model 3. A plot of the difference in the logarithm of the odds between assignment for Model 3 and assignment for Full Model 3 versus the logarithm of the odds of assignment for Model 3 was made. No discernible patterns could be seen, but it was interesting to notice that the majority of the values were less than zero. This can also be seen in Table 5.12. This indicates that the probability of assignment to Group 1 (the less seriously ill patients) is slightly higher for Full Model 3 than it is for Model 3.

† same definition in section 3.2.1

Table 5.12

Statistics for the difference in the logits for Model 3 and Full Model 3

MEAN	STDEV	MEDIAN	Q1	Q3	MIN	MAX
-0.126	0.174	-0.095	-0.198	-0.028	-1.24	0.368

Another comparison between the clinical classification and the Full Model 3 fit can be obtained from a comparison of the estimated parameters for Full Model 3 with their counterparts under the clinical classification. The estimated proportions in the two groups for Full Model 3 are $\hat{\pi}_1 = 0.5652$ and $\hat{\pi}_2 = 0.4348$. This is comparable to the sample proportions of 0.577 and 0.423 for the clinical classification. The parameter estimates for Full Model 3 are displayed in Table 5.13. It can be seen that these estimates are very similar to those displayed in Table 3.3.

Table 5.13

Summary statistics of the 12 pretreatment variables according to the Full Model 1 estimates.

(i) Continuous variables.

		Model Classification	
Variable	**Group**	**MEAN**	**STDEV**
Age	1	71.5	6.76
	2	71.3	7.65
HG	1	138.0	17.62
	2	129.9	20.82
sqrt(SZ)	1	2.9	1.23
	2	4.2	1.69
SG	1	8.9	1.15
	2	12.1	1.41
log(AP)	1	1.6	0.50
	2	3.9	1.68

The three variables in one partition have a mean vector of

$$\hat{\boldsymbol{\mu}}_{1\ell} = (\text{Wt}, \quad \text{SBP}, \quad \text{DBP}) = (100.4, \quad 14.5, \quad 8.3)$$

$$\hat{\boldsymbol{\mu}}_{2\ell} = (\text{Wt}, \quad \text{SBP}, \quad \text{DBP}) = (\ 97.2, \quad 14.2, \quad 8.0)$$

and variance covariance matrices of,

$$\hat{\Sigma}_{1\ell} = \begin{pmatrix} 173.60 & 6.86 & 4.60 \\ 6.86 & 6.43 & 2.44 \\ 4.60 & 2.44 & 2.42 \end{pmatrix}, \qquad \hat{\Sigma}_{2\ell} = \begin{pmatrix} 182.83 & 4.78 & 4.57 \\ 4.78 & 4.99 & 1.81 \\ 4.57 & 1.81 & 1.77 \end{pmatrix}$$

(ii) Categorical variables.

Level probabilities from the Model Classification using Model 1.

Variable	Group	Level probabilities						
PF	1	0.929	0.064	0.007	0.000			
	2	0.855	0.086	0.050	0.009			
HX	1	0.504	0.496					
	2	0.669	0.331					
EKG	1	0.335	0.045	0.075	0.060	0.311	0.170	0.004
	2	0.347	0.049	0.139	0.043	0.294	0.128	0.000
BM	1	0.993	0.007					
	2	0.634	0.366					

The program written for this thesis writes out the estimated values for the missing continuous data values for each group fitted in the model. The estimated values for the missing continuous variables in the cancer data set, are displayed in Table 5.14 for Full Model 3.

Table 5.14

Posterior probabilities of assignment to Group 1 and estimates of the missing continuous data values for Full Model 3.

Obs.	$\hat{\tau}_{i1}$	Missing Variable	Estimate $\hat{x}_{ij1}^{(t)}$	Estimate $\hat{x}_{ij2}^{(t)}$
2	0.373	10	8.93	12.08
42	0.000	1	71.54	71.31
48	0.991	9	2.94	4.19
57	0.365	10	8.93	12.08
74	1.000	9	2.94	4.19
123	0.925	10	8.93	12.08
125	0.971	10	8.93	12.08
131	1.000	9	2.94	4.19
158	0.004	9	2.94	4.19
169	0.000	10	8.93	12.08
193	0.741	4	102.24	99.86
262	0.006	4	97.88	94.66
336	0.981	10	8.93	12.08
418	0.926	10	8.93	12.08
436	0.927	10	8.93	12.08
471	0.985	4	100.43	97.20
471	0.985	5	14.50	14.17
471	0.985	6	8.28	7.98
471	0.985	8	137.97	129.95
473	0.847	4	100.43	97.20
473	0.847	5	14.50	14.17
473	0.847	6	8.28	7.98
473	0.847	8	137.97	129.95
474	0.669	10	8.93	12.08
475	0.910	4	100.43	97.20
475	0.910	5	14.50	14.17
475	0.910	6	8.28	7.98
475	0.910	8	137.97	129.95
475	0.910	9	2.94	4.19
475	0.910	10	8.93	12.08
484	0.144	10	8.93	12.08
485	0.000	9	2.94	4.19
488	0.000	4	100.43	97.20
488	0.000	5	14.50	14.17
488	0.000	6	8.28	7.98
488	0.000	8	137.97	129.95
488	0.000	9	2.94	4.19
506	0.000	10	8.93	12.08

Variables 4, 5, and 6 correspond to the variables Wt, SBP and DBP. These three variables are in one partition. It can be seen from Table 5.14 that both observations

193 and 262 had values recorded for the variables SBP and DBP, whilst the variable Wt was a missing data value. The techniques described in section 4.5.1 have been used to estimate the missing Wt values by computing the conditional mean of the data given the set of values observed for that observation and that partition.

It can be seen that the inclusion of the observations that had missing values in the twelve pre-trial variables, did not greatly affect the clustering. In the cancer data set, the percentage of data missing is actually fairly small. This is in contrast to the iris data sets analysed in sections 5.1 and 5.2, where there was quite an extreme amount of missing data. In section 5.1, the model has been able to detect the structure present in the data where the clusters have been well separated and there was an extreme amount of missing data. In section 5.2, it has been demonstrated that the model has performed well in detecting the structure in the overlapping clusters whilst simultaneously coping with the extreme amount of missing data.

It has been demonstrated that the approach of multivariate mixture model can cope with both missing categorical and missing continuous data values, as well as the hypothetical unobserved indicator variables of class membership. The variance covariance estimates are always reasonable and positive semi definite, unlike estimates calculated using all available pairs of variables of data.

Chapter 6

Summary and concluding remarks

6.1 Summary

Chapter 1 presents a review of the literature on the current methods of grouping data. A brief description of cluster analysis is given and some disadvantages of using cluster analysis as a method for grouping data are pointed out. A model based approach to clustering is examined.

Chapter 2 presents the detail for mixtures of multivariate normal distributions and latent class analysis. The assumption of the latent class model, namely the condition of local independence is discussed. An approach for handling data consisting of both categorical and continuous variables using mixtures of location models is given, and the general approach to multivariate mixture models is presented.

Chapter 3 demonstrates our ability to analyse data using the approach of multivariate mixture models by clustering several data sets. The ability of the method to handle large, complex data sets is illustrated by considering the clustering of the complete cases of the pre-trial variables of Byar and Green's (1980) Prostrate cancer data set. This consists of twelve variables measured on each of 475 observations. Eight of the variables could be taken as continuous variables, and four as categorical variables. To compare the approach of multivariate mixture models with that of Everitt and Mérette (1990), several models are fitted to Fisher's iris data. The extent of recovery of the structure known to exist in this data set is evaluated by calculating an adjusted rand index for each models fitted. In both of these data sets, the cluster assignments resulting from assigning each observation to the group of greatest probability, are compared with the grouping that is available – for the cancer data, the available grouping is the clinical classification of 'Stage', and for the iris data, the grouping is the species classification. In both examples, the model performed exceedingly well.

Data from a statistics examination are also clustered. Unlike the previous two data sets clustered, there is no *a priori* knowledge of a group structure for this data set. The models fitted using the approach of multivariate mixture models, appeared to cluster the observations in a meaningful way.

Chapter 4 introduces the problem of missing observations which occurs often in multivariate data sets. This chapter demonstates how our ability to use the approach of multivariate mixture models can be extended to include the facility to handle data sets where data are missing at random in the sense of Rubin (1976).

Chapter 5 uses the approach of mixture models to cluster multivariate data sets where data are missing at random. The approach is initially applied to a mixture of two multivariate normal distributions, where it is known that the two clusters are well separated. The approach is then applied to a mixture of two multivariate normal distributions, where it is known that the two clusters are poorly separated. In both cases, the data sets contained a fairly extreme amount of missing data. The extent of recovery of the true structure is then investigated and it is found that the approach of multivariate mixture models with data missing at random performed satisfactorily. The scope of the model is then illustrated by clustering the cases of the pre-trial variables of the Prostrate cancer data set, where observations with missing values in any of the pre-trial variables are included in the data set. The inclusion of these extra 31 observations only changes the cluster assignment of three observations, and two of these are not decisively assigned to a group under any of the models fitted.

The research reported in this thesis has shown that the approach of finite mixture models can easily cluster data sets containing both categorical and continuous variables.

6.2 Concluding remarks

For fully categorical datasets, the method of Latent Class Analysis has become a popular method of discovering underlying cluster structure. The class of models introduced and utilized in this thesis forms a natural extension of this class to data sets containing both categorical and continuous variables. Like Latent Class models, the models make free use of local independence to reduce the number of parameters in the model and to lead to descriptions of the clusters that can be easily understood. Provision is made, however, for the cautious introduction of within-cluster associations between variables.

Because the EM algorithm is used to fit the models, it is feasible to fit them to data sets with many variables and observations. As our ability to analyse data using multivariate mixture models includes the facility to handle situations where data are missing at random, fitting these models becomes an alternative to conventional cluster analysis algorithms. Missing data often presents problems for deterministic clustering algorithms.

The choice of discrete, multivariate normal and location model distributions as

the 'atoms' out of which our models are built has been made consciously in an effort to be bland and generic, but in situations where more is known about the nature of the distributions in subpopulations other types of distributions could be used in place of these.

Either taken as presented here, or modified to incorporate subject-area knowledge of distributions and parameters, multivariate mixture models should prove an invaluable tool in exploring large complex data sets.

There is scope for future research using multivariate mixture models. The standard errors of the estimates need to be investigated. It is possible that convergence of the EM algorithm could be sped up by the methods mentioned in section 1.4, although slowness of convergence was not a problem in the examples analysed in this thesis. The model could also be applied to different structured covariance matrices, for example those encountered in time series of measurements.

Appendix 1

Notes on Program Multimix

This program fits a mixture of multivariate distributions using the EM algorithm. The models that can be fitted are multivariate normal, latent class and location models having one categorical variable. The data file can contain both categorical and continuous data, or either of these data types. The NAG subroutines for matrix inversion and calculation of determinants are used in this program, hence the program needs linking into the NAG library.

Note: The desired form of the data matrix is to have variables in a partition cell being contiguous. To achieve this, the data is read in by specifying the column of the data array into which the J^{th} variable of the data file is stored in an order variable, JP(J). All further references to the variable J, refer to the rearranged order of the variables.

The program currently has a maximum of

1500 observations	(IOB = 1500)
6 groups	(IK6 = 6)
15 variables and partition cells	(IP15 = 15)
10 levels of categories	(IM10 = 10)
200 iterations to convergence	(ITER = 200)

NB. If these parameters are altered, remember to alter parameters (IK6 and IP15) in the subroutine, DETINV.

The parameter file contains:-

NG - the number of groups (distributions) in the finite mixture to be fitted.
NOBS - the number of observations.
NVAR - the number of variables.
NPAR - the number of partition cells (sets of variables associated within each distribution).
JP(J) - The column of the data array into which the J^{th} variable of the data file will be stored, J = 1, NVAR. For example, suppose that we want the 3rd variable in the first column, variable 4 in the second column, variable 7 in the 3rd column, and then variables 1, 2, 5 and 6. Then JP(J) = 4 5 1 2 6 7 3, for J = $1, \ldots, 7$.
IP(L) - the number of variables in the L^{th} partition cell, L = $1, \ldots,$ NPAR.

IPC(L) - number of continuous variables in the L^{th} partition cell.

ISV(L) - partition cell L starts at variable J.

e.g. if variables 6, 7, and 8, are in the same partition cell, then ISV(L) = 6, and IEV(L) = 8.

IEV(L) - partition cell L ends at variable J.

IPARTYPE(L) - indicator giving the type of model for each partition cell.

$$\text{IPARTYPE(L)} = \begin{cases} 1 & \text{for a categorical model;} \\ 2 & \text{for a multivariate normal model;} \\ 3 & \text{for a location model.} \end{cases}$$

IVARTYPE(J) - an indicator for the type of each variable

$$\text{IVARTYPE(J)} = \begin{cases} 1 & \text{for a categorical variable;} \\ 2 & \text{for a multivariate normal variable;} \\ 3 & \text{for a categorical variable in a location model;} \\ 4 & \text{for a multivariate normal variable in a location model,} \end{cases}$$

NCAT(J) - the number of categories for the J^{th} categorical variable. For continuous variables, NCAT(J) is entered as 0.

ISPEC - indicator variable determining whether the observations are specified into groups.

$$\text{ISPEC} = \begin{cases} 1 & \text{observations are not specified into groups;} \\ 2 & \text{observations are specified into groups.} \end{cases}$$

(1) ISPEC = 1 — read in the estimates of the parameters.

PI(K)- estimated mixing proportions for each group.

THETA(K,J,M) - estimated probability that the J^{th} categorical variable is at level M, given that in group K

EMU(K,L,J) - estimated mean vector for group K, partition cell L and variable J.

EMUL(K,L,J,M) - estimated mean vector for group K, partition cell L, variable J, at the M th level of the categorical variable in the location model.

VARIX(K,L,I,J) - estimated covariance between variables I and J for group K, partition cell L where $\text{I} = 1, \ldots, \text{IPC(L)}$ and $\text{J} = 1, \ldots, \text{IPC(L)}$.

Note: The parameters that are required, are read in for each partition cell, $\text{L} = 1, \ldots, \text{NPAR}$. For example, if the variables within the partition cell are all categorical, that is, ITYPE(L) = 1, then THETA(K, J, M), for $\text{M} = 1, \ldots, \text{NCAT(J)}$ is required for the variable in that partition cell.

If the variables within the partition cell are continuous, multivariate normal variables, that is, ITYPE(L) = 2, then estimates of EMU(K,L,J) are required for each variable.

If the variables within the partition cell follow the location model, that is, ITYPE(L) = 3, then $\text{THETA(K, J, M)}, \text{M} = 1, \ldots, \text{NCAT(J)}$ is required for the

categorical variable, and EMUL(K, L, J, M), M = 1, . . . , IM(L) is required for each continuous multivariate normal variable. (Note that IM(L) is the number of categories of the categorical variable associated with the location model.)

The estimates are read in for group 1, and then for group 2 etc.

(2) ISPEC = 2, read in

IGRP(I) - variable specifying which group each observation is in. After reading in this variable, the program proceeds to the M step to calculate estimates of the parameters.

Creating an input file

A FORTRAN program has been written to help set up the parameter input file for program MULTIMIX. If ISPEC = 2, that is, the grouping of the data is specified, the program requires the grouping to be in a separate file. This file must be in existence before the program for creating the parameter input file is run. This program has a very basic error subroutine to check whether the number of variables in each partition cell matches with the total number of variables, and the type of each variable matches with the partition cell type. This program is being extended to facilitate ease of input.

Format of the parameter file

Format is free field.
Input:-
NG NOBS NVAR NPAR ISPEC
(JP(J), J = 1, NVAR)
(IP(L), L = 1, NPAR)
(IPC(L), L = 1, NPAR)
(ISV(L), L = 1, NPAR
(IEV(L), L = 1, NPAR)
(IPARTYPE(L), L = 1, NPAR)
(IVARTYPE(J), J = 1, NVAR)
(NCAT(J), J = 1, NVAR)

If ISPEC = 1, read in estimates of the parameters. See section below.
If ISPEC = 2, read in specified grouping of observations
(IGRP(I), I = 1, NOBS)

Estimates of the parameters

(PI(K), K = 1, NG)
****
For each partition cell, L = 1, NPAR read in the required parameters.
(THETA(K,J,M), M = 1, NCAT(J)) - for categorical variables
- repeat for each variable, J = ISV(L), IEV(L)
(EMU(K,L,J), J = 1, IPC(L)) - for the multivariate normal model
(THETA(K,J,M), M = 1, NCAT(J)) - for the categorical variable in location model
(EMUL(K,L,J,M), M = 1, IM(L)) for each continuous variable J = 1, IPC(L) in the
location model

Repeat from **** for each group.
Read in the estimates of the variance for each partition cell. (For continuous variables only)
((VARIX(K,L,I,J), J = 1, IPC(L)), I = 1, IPC(L)) - repeat for each group.

Appendix 2

Statistics Examination paper

DEPARTMENT OF MATHEMATICS AND STATISTICS

Statistical Methods

Name: _____ ID Number: _____

SECTION A: Multiple Choice

Answer by making a mark in the appropriate square

1. The **standard error** of an estimator is

[ ] the population standard deviation.
[ ] the standard deviation of its sampling distribution.
[ ] the difference between the estimator and the parameter.
[ ] the bias of the estimator.
[ ] none of these, it's actually ...

2. Only one of the following statements about confidence intervals is always true. Which one?

[ ] they take the form *estimate plus-or-minus multiple of standard error.*
[ ] the parameter must lie within the confidence interval.
[ ] the interval must be finite in length.
[ ] the interval contains the parameter with probability equal to the confidence coefficient.

3. The **significance level**, α, of a hypothesis test or significance test is

[ ] the probability that the alternative hypothesis is false.
[ ] the probability that the null hypothesis is true.
[ ] the probability of rejecting the null hypothesis when it is true.
[ ] how seriously a result should be regarded.
[ ] none of these, in fact it is ...

4. The **power function** of a hypothesis test is

[ ] the probability of rejecting Ho as a function of the true parameter value.
[ ] one minus the significance level.
[ ] a Minitab subcommand used for exponents.
[ ] another name for the type II error.
[ ] none of these, in fact it is ...

5. The **P-value** of a significance test is

[ ] the true value of a population proportion.
[ ] the sample value of the proportion.
[ ] the smallest value of a at which the null hypothesis would be rejected.
[ ] the size of the critical region.
[ ] none of these, it's actually ...

6. Suppose that we have a $100(1-\alpha)\%$ confidence interval for a parameter θ, then we can test the hypothesis $H_0 : \theta = \theta_0$ at significance level α against the alternative hypothesis $H_1 : \theta \neq \theta_0$ by

[ ] rejecting H_0 whenever θ_0 is inside the confidence interval.
[ ] rejecting H_0 whenever θ_0 is outside the confidence interval.
[ ] rejecting H_0 whenever $\bar{x}$ is inside the confidence interval.
[ ] rejecting H_0 whenever $\bar{x}$ is outside the confidence interval.
[ ] none of these, the way to do it is ...

7. A collection, or set, of individuals, objects, or measurements whose properties are to be analysed is

[ ] a parameter.
[ ] a population.
[ ] a statistic.
[ ] a variable.
[ ] an experiment.

8. When testing the hypothesis $H_0 : \sigma^2 = 100$ against the alternative $\sigma^2 < 100$ using a sample of size 15 and a significance level of .05, what number forms the boundary of the critical region.

[ ] $t\,(14, .05)$.
[ ] $\chi^2(15, .05)$.
[ ] $z\,(.95)$.
[ ] 6.571.
[ ] 23.685.

9. The sign test is a useful non-parametric method that can be used for a hypothesis test concerning

[ ] the difference between two independent means.
[ ] the value of the median for one population.
[ ] the spread of one population.
[ ] the variances of two related samples.
[ ] none of these, you use it to ...

SECTION B : Answer on the paper in the space provided.

10. When a fair coin is tossed, suppose that a 'tail' scores 2 and a 'head' scores 4. Calculate the probability distribution function, its mean, and its variance, for the following random variables:

(a) The score on a single toss

(b) The mean score on two independent tosses.

Does the relationship between the variances remind you of a general rule? State this rule.

11. An electronic method has been developed for measuring the percentage of fat in cheese. If the true percentage of fat is θ, the method gives a result X normally distributed with mean 1.050 and variance 0.4.

(a) If we follow the electronic procedure 3 times independently, obtaining the results 29.3, 30.0, 29.5, give a 95% confidence interval for θ.

(b) How many electronic fat determinations would we need to do, to obtain a 95% confidence interval of half-width 0.01?

12. Suppose that you have a summer job at Horotiu Freezing Works. Your boss was pleased with the help you gave him earlier and seeks your aid a second time. Another small fire has occurred and he wishes to recheck that a batch containing a large number of carcasses is OK. This time he has already decided to use an (n, c) plan with $n = 15$ and $c = 1$. He wants you to tell him the probability of rejecting a batch with (a) 50% (b) 20% and (c) 5% of defective (ie smoke-damaged) carcasses. Work out these probabilities.

13. The following data consists of weight gains in kilograms for ten pairs of identical twins in a designed experiment to evaluate two systems of feeding. One calf in each pair is randomly assigned to be fed by System A, and the other is fed by System B.

	Pair									
System	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	19.0	17.7	17.6	19.1	20.9	19.5	17.2	19.5	23.2	19.5
B	18.3	17.3	16.8	20.0	19.1	18.2	17.3	20.0	23.0	17.8

(i) Describe two alternative tests which could be used to test the hypothesis that the weight gains under the two systems are the same. For each test state the test statistic, what distribution it has when the null hypothesis is true, and how the critical region is constructed. Do not carry out the numerical calculations.

(ii) Analyse the data by carrying out one of the two procedures described in part (i). Use either $\alpha = 0.05$ or a P-value approach.

References

Afifi, A. A. and Elashoff, R. M. (1966). Missing observations in multivariate statistics I: Review of the literature, *J. Am. Statis. Assoc.* **61**, 595–604.

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical Modelling of Data on Teaching Styles. *J. R. Statist. Soc.* A, **144**, 419–461.

Aitkin, M. and Rubin, D. B. (1985). Estimation and Hypothesis Testing in Finite Mixture Models. *J.R. Statist. Soc.* B, , **47**, 67–75.

Aitkin, M. and Tunnicliffe Wilson, G. (1980). Mixture models, Outliers and the EM Algorithm. *Technometrics*, **22**, 325–331.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Automat. Contr.*, **AC–19**, 716–723.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*, New York : Wiley.

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

Bain, L. J. and Englehardt, M. (1992). *Introduction to probability and mathematical statistics*, 2nd ed., PWS–KENT, Boston.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis, *Journal of the Royal Statistical Society, Series* B, **37**, 129–145.

Beaton, A. E. (1964). The use of special matrix operators in statistical calculus. Educational Testing Service Research Bulletin, RB–64–51.

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

Bozdogan, H. and Sclove, S. L. (1984). Multi-sample cluster analysis using Akaike's information Criterion. *Ann. Inst. Statist. Math.*, **36**, 163–180.

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

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

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

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

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

Dempster, A. P. (1969). *Elements of continuous multivariate analysis.* Reading, MA: Addison-Wesley.

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

Dempster, A. P., and Rubin, D. B. (1983). Overview, in *Incomplete data in sample surveys, Vol. II: Theory and annotated bibliography* (W. G. Madow, I. Olkin, and D. B. Rubin, Eds.). New York : Academic Press, 3–10.

Everitt, B. S. (1980). *Cluster Analysis.* (2nd edition), Gower Publications, London.

Everitt, B. S. (1981). A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioural Research*, **16**, 171–180.

Everitt, B. S. (1984a). *An introduction to latent variable models.* Chapman and Hall, London.

Everitt, B. S. (1984b). A note on parameter estimation for Lazarsfelds latent class model using the EM algorithm. *Multivariate Behavioural Research*, **19**, 79–89.

Everitt, B. S. (1984c). Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions; a comparison of different algorithms. *The Statistician*, **33**, 205–215.

Everitt, B. S. (1985). Mixture Distributions. In *Encyclopedia of Statistical Sciences*,(vol. 5), S. Kotz and N.L. Johnson (Eds.). New York: Wiley, 559–569.

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

Everitt, B. S. and Dunn, G. (1991). *Applied Multivariate Data Analysis.* London : Edward Arnold.

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

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

Feng, Z. D. and McCulloch, C. E. (1994). On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variances. *Biometrics* **50**, 1158–1162.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–184.

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

Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio test statistic for the mixure model and related results. Proc. Berkeley Conference in Honor of Jerzy Newman and Jack Kiefer (Vol. II), L.M. Le Cam and R.A. Olshen (Eds.). Monterey: Wadsworth, 789–806.

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

Goodnight, J. H. (1979). A tutorial on the sweep operator. *American Statistician*, **33**, 149–158.

Gordon, A. D. (1981). *Classification.* London: Chapman and Hall.

Hand, D. J. (1981). *Discrimination and Classification*, John Wiley and Sons, London.

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

Hartigan, J. A. (1977). Distribution problems in clustering. In *Classification and Clustering*, J. Van Ryzin (Ed.). New York: Academic Press, 45–71.

Hartigan, J. A. (1985a). A failure of likelihood asmptotics for normal mixtures. Proc. Berkeley Conference in Honor of Jerzy Newman and Jack Kiefer (Vol. II), L. M. Le Cam and R. A. Olshen (Eds.). Monterey: Wadsworth, 807–810.

Hartigan, J. A. (1985b). Statistical theory in clustering. *J. Classification*, **2**, 63–76.

Hartley, H. O. and Hocking, R. R. (1971). The analysis of incomplete data, *Biometrics* **27**, 783–808.

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

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

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.*, **13**, 795–800.

Heiberger, R. M., (1989). *Computation for the analysis of designed experiments*, New York : Wiley.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classification*, **2**, 193–218.

James, M. (1985). *Classification Algorithms.* Collins, London.

Jardine, N. and Sibson R. (1971). Mathematical taxonomy London, New York, Wiley.

John, S. (1970). On identifying the population of origin of each observation in a mixture of observations from two normal populations. *Technometrics*, **12**, 553–563.

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

Krzanowski, W. J. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data, *Comp. and Maths. with Appls.*, **12A(2)**, 179–185.

Krzanowski, W. J. (1988). *Principles of multivariate analysis : A user's perspective.* Oxford University Press, New York.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems *Computer J.*, **9**, 373–380.

Lauritzen, S. L., and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, **17**, 31–57.

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

Lehmann, E. L. (1980). Efficient likelihood estimators. *Amer. Statistician* **34**, 233–235.

Little, R. J. A. (1982). Models for nonresponse in sample surveys, *J. Am. Statis. Assoc.* **77**, 237–250.

Little, R. J. A. and Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete data likelihood, *American Statistician*, **37**, 218–220.

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

Little, R. J. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values, *Biometrika*, **72**, 497–512.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society, Series* B, **44**, 226–233.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis.* New York: Academic Press.

Marriott, F. H. C. (1971). Practical problems in a method of cluster analysis. *Biometrika*, **27**, 501–515.

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

Mc Hugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, **21**, 331–347.

McLachlan, G. J. (1987). On Bootstrapping the Likelihood Ratio Test Statistic for the number of components in a Normal Mixture. *Journal of Applied Statistics*, **36**, 318–324.

McLachlan, G. J. (1988). On the choice of starting values for the EM algorithm in fitting mixture models. *The Statistician*, **37**, 417–435.

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

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

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms, *Journal of the Royal Statistical Society, Series* B, **59**, 127–138.

Milligan, G. W. and Cooper, M. C. (1986). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation. *Computer Journal*, **20**, 359–363.

Morrison, D. F. (1976). *Multivariate statistical methods.* 2nd ed. New York : McGraw-Hill.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables, *Ann. Math. Statist.*, **32**, 448–465.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle : theory and applications, *Proceedings of the* 6^{th} *Berkeley Symposium on Mathematical Statistics and Probability.* **1**, 697–715.

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

Quinn, B. G., McLachlan, G. J. and Hjorth, N. L. (1987). A note on the Aitken–Rubin approach to hypothesis testing in mixture models. *Journal of the Royal Statistical Society, Series B*, **49**, 311–314.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, **26**, 295–239.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–593.

Rubin, D. B. (1991). EM and beyond. *Psychometrika*, **56**, No. 2, 241–254.

Rubin, D. B. and Szatrowski, T. H. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika*, **69**, 3, 657–660.

Santos, R. (1981). Effects on imputation on regression coefficients. In *Proc. Survey Res. Meth. Sect., Am. Statist. Assoc.*, 140–145.

Schervish, M. J. (1984) Algorithm AS195. Multivariate normal probabilites with error bound. *J.R. Statist. Soc.* C, **33**, 81–87.

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

Sclove, S. L. (1977). Population mixture models and clustering algorithms. *Commun. Statist.–Theor. Meth.*, **A6**, 417–434.

Sclove, S. L. (1983). Application of the conditional population-mixture model to image segmentation. *IEEE Trans. Pattern Anal. Machine Intelligence.* **PAMI-5**, 428–433.

Seber, G. A. F. (1977). *Linear Regression Analysis*, New York : Wiley.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, **1**, 49–58.

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

Tan, W. Y. and Chan, W. C. (1972). 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.

Titterington, D. M. (1981). Contribution to the discussion of paper by M. Aitkin, D. Anderson and J. Hinde. *J.R.Statist. Soc.* A **144**, 459.

Titterington, D. M. (1984). Comments on a paper by S.L.Sclove. *IEEE Trans. Pattern Anal. Machine Intelligence.* **PAMI-6**, 656–658.

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

Uebersax, J. S. and Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, **9**, 559–572.

Wolfe, J. H. (1967). NORMIX: Computational methods for estimating the parameters of multivariate normal mixtures of distributions. Research memo. SRM **68–2**. San Diego: U.S. Naval Personnel and Training Research Laboratory.

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

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

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, **11**, 95–103.

Zhaorong, J., McGilchrist, C. A. and Jorgensen, M. A. (1992). Mixed model discrete regression. *Biom. J.*, **34**, 691–700.