ELSEVIER Computational Statistics & Data Analysis 41 (2003) 429–440

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Mixture model clustering for mixed data with missing information

Lynette Hunt*, Murray Jorgensen

Department of Statistics, University of Waikato, Hamilton, New Zealand

Received 1 March 2002

Abstract

One difficulty with classification studies is unobserved or missing observations that often occur in multivariate datasets. The mixture likelihood approach to clustering has been well developed and is much used, particularly for mixtures where the component distributions are multivariate normal. It is shown that this approach can be extended to analyse data with mixed categorical and continuous attributes and where some of the data are missing at random in the sense of Little and Rubin (Statistical Analysis with Mixing Data, Wiley, New York).
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Clustering; Mixed data; Missing at random

1. Introduction

Missing observations are frequently seen in multivariate data sets. For example, the specimen may be damaged and thus not all attributes can be measured, or an inexpensive and easy administered test may be administered to all items in the sample whilst the more expensive test may only be administered to a random sub-sample of the items. In such situations, the data matrix will be incomplete with not all attributes being observed for all items. These missing values can be regarded as accidental missing values.

Review papers in the literature on partially missing data include those by Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), and

* Corresponding author. Fax: +64-7-838-4155.
E-mail address: lah@stats.waikato.ac.nz (L. Hunt).
URL: http://www.stats.waikato.ac.nz/Staff/index.html

PII: S0167-9473(02)00190-1

Dempster et al. (1977), and monographs on partially missing data by Little and Rubin (1987), and Schafer (1997). The approaches appropriate for handling such data in classification studies are restricted due to the reluctance of the investigator to make assumptions about the data (Gordon, 1999) and the lack of a formal model for cluster analysis. Given the objective of clustering the data, we need to implement some technique when the data to be clustered are incomplete.

Gordon (1999, p. 26) notes that Gower's (1971) general (dis)similarity coefficient can be used as one strategy to cope with missing variables, by assuming that the contribution that would have been provided by the incompletely recorded variable to the proximity between the two items is equal to the weighted mean of the contributions provided by the variables for which complete information is available.

Data are described as 'missing at random' when the probability that a variable is missing for a particular individual may depend on the values of the observed variables for that individual, but not on the value of the missing variable. That is, the distribution of the missing data mechanism does not depend on the missing values. For example, censored data are certainly *not* missing at random.

Rubin (1976) showed that the process that causes the missing data can be ignored when making likelihood-based inferences about the parameter of the data if the data are 'missing at random' and the parameter of the missing data process is 'distinct' from the parameter of the data. When the data are missing in this manner, the appropriate likelihood is simply the density of the observed data, regarded as a function of the parameters. 'Missing at random' is a central concept in the work of Little and Rubin (1987).

The EM algorithm of Dempster et al. (1977) is a general iterative procedure for maximum likelihood estimation in incomplete data problems. Their general model includes both the conceptual missing data formulation used in finite mixture models and the accidental missing data discussed earlier. Many authors, for example McLachlan and Krishnan (1997), have discussed the EM algorithm and its properties.

Little and Schluchter (1985) present maximum likelihood procedures using the EM algorithm for the general location model with missing data. They note that their model reduces to that of Day (1969) for K-variate mixtures when there is one K-level categorical variable that is completely missing. Little and Rubin (1987) and Schafer (1997) point out that the parametric mixture models lend themselves well to implementing incomplete data methods. We implement their approach to produce explicit methodology that enables the clustering of mixed (categorical/continuous) data using a mixture likelihood approach when data are missing at random. We illustrate this approach by clustering Byar's prostate cancer data. It is shown that the proposed methodology can detect meaningful structure in mixed data when there is a fairly extreme amount of missing information.

2. The mixture approach to clustering data

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

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

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

The EM algorithm of Dempster et al. (1977) is applied to the finite mixture model by viewing the data as incomplete. In the case of mixtures of distributions, the ‘missing’ data are the unobserved indicators of group membership. Let the vector of indicator variables, $\mathbf{z}_i=(z_{i1},\dots,z_{iK})'$, be defined by

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

where $\mathbf{z}_i, i=1,\dots,n$, are independently and identically distributed according to a multinomial distribution generated by a single trial of an experiment with K mutually exclusive outcomes having probabilities $\pi_1,\dots,\pi_K$.

Let $\hat{\boldsymbol{\phi}}$ denote 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 estimated posterior probability that observation $\mathbf{x}_i$, belongs to group k, is given by

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

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

$$\hat{z}_{ik}>\hat{z}_{ik'}\quad\text{for } k=1,\dots,K,\quad k\neq k'.$$

Finite mixture models are frequently fitted where the component densities $f_k(\mathbf{x};\boldsymbol{\theta}_k)$ are taken to be multivariate normal; i.e., $\mathbf{x}_i\sim N_p(\boldsymbol{\mu}_k,\Sigma_k)$, if observation i belongs to group k. This model has been studied by Titterington et al. (1985), and by McLachlan and Basford (1988). Further details on the maximum likelihood estimates of the components of $\boldsymbol{\phi}$ can be found in McLachlan and Peel (2000, p. 82).

The latent class model described, for example, by Everitt (1984), is a finite mixture model for data where each of the p attributes is discrete. Suppose that the jth attribute can take on levels $1,\dots,M_j$ and let λ_{kjm} be the probability that for individuals from group k, the jth attribute has level m. Then, conditional on individual i belonging to group k, $f_k(\mathbf{x}_i;\theta_k)=\prod_{j=1}^{p}\lambda_{kjx_{ij}}$. In other words, within each group the distributions of the p attributes are independent. This property has been termed *local independence*.

2.1. Multimix

Jorgensen and Hunt (1996) and Hunt and Jorgensen (1999) proposed a general class of mixture models to include data having both continuous and categorical attributes.

This model, which they dubbed the ‘Multimix’ model, was conceived of initially as a joint generalization of both latent class models and mixtures of multivariate normal distributions. They suggested an approach based on a form of local independence by partitioning the observational vector $\mathbf{x}_i$ such that

$$\mathbf{x}_i = (\breve{\mathbf{x}}_{i1}|\ldots|\breve{\mathbf{x}}_{il}|\ldots|\breve{\mathbf{x}}_{iL})',$$

where the attributes within partition cell $\breve{\mathbf{x}}_{il}$, are independent of the attributes in partition cell $\breve{\mathbf{x}}_{il'}$, for $l \neq l'$ within each of the K sub-populations. Thus if individual i belongs to group k, we can write

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

In this paper, we restrict ourselves to the following distributions suggested for the partition cells:

Discrete distribution: where $\breve{\mathbf{x}}_{il}$ is a one-dimensional discrete attribute taking values $1,\ldots,M_l$ with probabilities λ_{klM_l}. We will denote this distribution by $D(\lambda_{kl1},\ldots,\lambda_{klM_l})$.

Multivariate Normal distribution: where $\breve{\mathbf{x}}_{il}$ is a p_l-dimensional vector with a $N_{p_l}(\boldsymbol{\mu}_{kl},\Sigma_{kl})$ distribution if individual i is in group k.

See Hunt and Jorgensen (1999) for the maximum likelihood estimates for the components of $\boldsymbol{\phi}$. This approach included the latent class model and mixtures of multivariate normal distributions as special cases.

2.2. Graphical models

A revealing alternative way of looking at these multivariate models is within the framework of graphical models described by Lauritzen and Wermuth (1989). In this framework graphs are associated with models. The graph of a model contains a vertex corresponding to each variable in the model. Edges are assigned such that the absence of a connected path between two vertices corresponds to independence of the corresponding variables. If no path exists between two vertices after a set of vertices has been removed, then this means the variables represented by the two vertices are independent *conditionally* on knowing the values of the variables corresponding to the removed vertices. Latent class models for p variables are represented by a graph on $p+1$ vertices corresponding to the p variables plus one categorical variable indicating the cluster. Each of the p variables are joined by a single edge to the cluster variable, and these are the only edges in the graph.

A *clique* in a graph is a maximal set of vertices such that an edge connects each pair of vertices in the set. No independence assumptions are made about variables in a clique. The graph of a *Multimix* model may be described as follows: corresponding to each cell in the partition of attributes there is a clique of vertices, each clique forms a complete graph, none of these graphs are directly connected, but all vertices in each are joined to an additional vertex that represents a categorical latent variable giving the cluster assignment of each observation. As a special case, locally independent multivariate mixture models may also be described in the language of graphical models by a graph in which the edges connect the latent cluster variable to each of the p manifest variables. If all variables in this special case are discrete we have a latent class model. Edwards (1995) provides a gentle introduction to the concepts of graphical modelling.

Multimix models with only continuous variables are mixtures of multivariate normals in which the covariance matrices are each block-diagonal with the same block pattern. Banfield and Raftery (1993) consider other kinds of restrictions to covariance matrices in mixtures of multivariate normals, with possible limitations on volume, orientation and shape of the component distributions.

2.3. Missing data

Little and Rubin (1987, Chapter 3) review several 'quick' methods for coping with missing data in multivariate statistical analyses. Essentially, they consider

(1) 'complete-case' methods, which discard observations in which any variable is missing;
(2) 'available-case' methods based on pairwise sample covariances using all observations in which both variables are observed;
(3) methods based on filling-in or 'imputing' the missing values.

They conclude '... it is hard to recommend any of the simple methods discussed since (1) their performance is unreliable; (2) they often require ad hoc adjustments to yield satisfactory estimates, and (3) it is not easy to distinguish situations when the methods work from when they fail'. Little and Rubin (1987) go on to develop methods for handling missing data based on the *EM* algorithm. Essentially their methods are of two kinds: those for which the missing data mechanism is *ignorable*, the data being missing at random, and those for which a model must be specified describing the mechanism by which the data come to be missing.

In this paper, we put forward a method for mixture model clustering based on the assumption that the data are missing at random and hence the missing data mechanism is ignorable. It is natural to ask whether we can do without the 'missing at random' assumption and work with non-ignorable missing data mechanisms. However, the effective use of non-ignorable models requires knowledge of the missing data mechanism that will quite often be lacking in an exploratory clustering situation. Molenberghs et al. (1999) discuss the use of non-ignorable models in the context of longitudinal categorical data. They note that such models cannot be tested by the data and advocate using a range of models in a sensitivity analysis, while employing as much context-derived information as possible. Because our interest in this paper is directed towards the general clustering problem, we confine ourselves to methods that are technically valid when the missing data mechanism is ignorable.

We now present a form of Multimix suitable for multivariate data sets with missing data. This model reduces to that given by Hunt and Jorgensen (1999) when all the data are observed.

Suppose we write the observation vector $\mathbf{x}_i$ in the form $(\mathbf{x}_{\mathrm{obs},i}, \mathbf{x}_{\mathrm{miss},i})$, where $\mathbf{x}_{\mathrm{obs},i}$ and $\mathbf{x}_{\mathrm{miss},i}$, respectively, denote the observed and missing attributes for observation i. This is a formal notation only and does not imply that the data are rearranged to achieve this pattern. In fitting the mixture model, there are now 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. However, these unintended missing values can also be of two different types. They may be continuous and belong to a multivariate normal partition cell, or a categorical variable involved in a partition cell with a discrete distribution.

The E step of the EM algorithm requires the calculation of $Q(\boldsymbol{\phi},\boldsymbol{\phi}^{(t)}) = E\{L_C(\boldsymbol{\phi})|\mathbf{x}_{\text{obs}};\boldsymbol{\phi}^{(t)}\}$, the expectation of the complete data log-likelihood 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

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

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

The remaining calculations in the E step require the calculation of the expected value of the complete data sufficient statistics for each partition cell l, conditional on the observed data and the current values of the parameters for that partition cell.

For each discrete partition cell l and each value m_l of $\breve{\mathbf{x}}_{il}$, the E step calculates

$$\begin{aligned}E(z_{ik}\delta_{ilm}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)}) &= \begin{cases}\hat{z}_{ik}\delta_{ilm}, & x_{il}\text{ observed}\\ \hat{z}_{ik}E(\delta_{ilm}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)}), & x_{il}\text{ missing}\end{cases}\\ &= \begin{cases}\hat{z}_{ik}\delta_{ilm}, & x_{il}\text{ observed,}\\ \hat{z}_{ik}\hat{\lambda}_{ilm}^{(t)}, & x_{il}\text{ missing,}\end{cases}\end{aligned}$$

where we have defined an indicator variable

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

Let

$$\hat{\delta}_{ilm} = \begin{cases}\delta_{ilm}, & x_{il}\text{ observed,}\\ \lambda_{ilm}, & x_{il}\text{ missing.}\end{cases}$$

Then this expectation can be written in the form

$$E(z_{ik}\delta_{ilm}|\mathbf{x}_{\text{obs},i};\theta k^{(t)}) = \hat{z}_{ik}\hat{\delta}_{ilm}$$

for $k = 1,\dots,K$; each categorical $\breve{\mathbf{x}}_{il}$ and each value m_l of $\breve{\mathbf{x}}_{il}$.

For multivariate normal partition cells, depending on the attributes observed for individual i in the cell, these expectations may require the use of the sweep operator described originally by Beaton (1964). The version of sweep we use is the one defined by Dempster (1969); also described in Little and Rubin (1987, pp. 112–119). Little and Rubin (1987) and Schafer (1997) demonstrate the usefulness of sweep in maximum likelihood estimation for multivariate missing-data problems. Hunt (1996) implemented this approach with mixtures of multivariate normal distributions. The approach is adapted in the following manner:

Suppose we form the augmented covariance matrix A_l using the current estimates of the parameters for group k in cell l where

$$A_l = \begin{pmatrix} -1 & \mu_{k1} & \mu_{k2} & \cdots & \mu_{kp_l} \\ \mu_{k1} & \sigma_{k11} & \sigma_{k12} & \cdots & \sigma_{k1p_l} \\ \mu_{k2} & \sigma_{k21} & \cdots & \cdots & \sigma_{k2p_l} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{kp_l} & \sigma_{kp1} & \cdots & \cdots & \sigma_{kp_lp_l} \end{pmatrix},$$

where the rows and columns of A_l are indexed from 0 to p_l. Then sweeping on row and column 1 corresponds to sweeping on x_{i1}, and sweeping on row and column j corresponds to sweeping on x_{ij}. Sweeping on the elements of A_l corresponding to the observed x_{ij} in cell l, yields the conditional distribution (regression) of the missing $x_{ij'}$ on the observed x_{ij} in the cell.

The remaining calculations in the E step for multivariate normal partition cells are as follows:

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

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

For $j \neq j'$,

$$\begin{aligned} &E(z_{ik}x_{ij}x_{ij'}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)}) \\ &\quad = \begin{cases} \hat{z}_{ik}x_{ij}x_{ij'}, & x_{ij} \text{ and } x_{ij'} \text{ observed}, \\ \hat{z}_{ik}x_{ij}E(x_{ij'}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)}), & x_{ij} \text{ observed, } x_{ij'} \text{ missing}, \\ \hat{z}_{ik}E(x_{ij}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)})x_{ij'}, & x_{ij} \text{ missing, } x_{ij'} \text{ observed}, \\ \hat{z}_{ik}[E(x_{ij}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)})E(x_{ij'}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)}) & \\ \quad + \text{Cov}(x_{ij},x_{ij'}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)})], & x_{ij} \text{ and } x_{ij'} \text{ missing}, \end{cases} \end{aligned}$$

for $i = 1,\ldots,n$; $k = 1,\ldots,K$; $x_{ij} \in \breve{\mathbf{x}}_{il}$ where $\breve{\mathbf{x}}_{il}$ is a multivariate normal partition cell.

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 $\mathbf{x}_{\text{obs},i}$ observed for that individual in that cell and the current estimates of the parameters for the cell. However, for the conditional expectations to be used in the calculation of the covariance matrix, i.e. $E(z_{ik}x_{ij}^2|\mathbf{x}_{\text{obs},i};\theta_k^{(t)})$ and $E(z_{ik}x_{ij}x_{ij'}|\mathbf{x}_{\text{obs},i};\theta_k^{(t)})$, then, respectively, if x_{ij} is missing, or if x_{ij} and $x_{ij'}$ are missing in that cell, the conditional mean of x_{ij} is adjusted by the conditional covariances as shown above. These conditional means and the non-zero conditional covariances are found by using the sweep operator on the augmented covariance matrix that has been created using the current estimates of the parameters for that particular multivariate normal partition cell. The augmented covariance matrix is swept on the observed attributes $\mathbf{x}_{\text{obs},i}$ in cell l such that these attributes are the predictors in the regression equation and the remaining attributes are the outcome variables for that cell.

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

Mixing proportions:

$$\hat{\pi}_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ik}^{(t)} \quad \text{for } k = 1,\ldots,K.$$

Discrete distribution parameters:

$$\hat{\lambda}_{klm} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{z}_{ik}\hat{\delta}_{ilm} \quad \text{for } k = 1,\ldots,K, \quad m = 1,\ldots,M_l$$

and where l indexes a discrete partition cell $\breve{\mathbf{x}}_l$.

Multivariate Normal parameters:

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

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

for $k=1,\ldots,K$. Here j and j' index the continuous attributes belonging to a multivariate normal cell $\breve{\mathbf{x}}_l$.

Let the conditional covariance between attributes j and j' for individual i, given that individual i belongs in group k,

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

and let the imputed value for attribute j of individual i, given the current value of the parameters and that the individual belongs in group k, be

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

The parameter estimates for the mean and the variance or covariance terms can be written in the form

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

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

for $k = 1,\ldots,K$. Here again j and j' index the continuous attributes belonging to a Multivariate Normal cell $\breve{\mathbf{x}}_l$.

3. Application

The approach will be illustrated by considering the clustering of cases on the basis of the pre-trial variables of the prostate cancer clinical trial data of Byar and Green (1980), reproduced in Andrews and Herzberg (1985, pp. 261–247). The data are available at http://lib.stat.cmu.edu/datasets/Andrews/T46.1. The data were obtained from a randomized clinical trial comparing four treatments for 506 patients with prostatic cancer. These patients had been grouped on clinical criteria into Stage 3 and Stage 4 of the disease. As reported by Byar and Green, Stage 3 represents local extension of the disease with no clinical evidence of distant metastasis, whilst Stage 4 represents distant metastasis as evidenced by acid phosphatase levels, X-rays, or both.

There are 12 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 treat SG as a continuous variable. Two of the discrete covariates have two levels, one has four levels and the fourth discrete covariate has seven levels. As detailed in Hunt and Jorgensen (1999), two variables, SZ and AP have been transformed to make their distributions more symmetric.

Thirty-one individuals have at least one of the pre-trial covariates missing, giving a total of 62 missing values. As only approximately 1% of the data are missing, more missing observations were created, where the probability of an observation on an attribute being missing was taken independently of all other data values. Missing values generated in this manner are missing completely at random, and the missing data mechanism is ignorable for likelihood inferences (Little and Rubin, 1987; Schafer, 1997).

Missing values were created by assigning each attribute of each individual a random digit generated from the discrete $[0,1]$ distribution, where the probability of a zero was taken, respectively, as 0.10, 0.15, 0.20, 0.25 and 0.30. Attributes for an individual were recorded as missing when the assigned random digit was zero. This process was repeated 10 times for each of the probabilities chosen.

We report fully the results taken from one pattern of missing data where the probability of an observation on an attribute being missing was 0.30. This illustrates the approach on a fairly extreme case of the type of data that would be analysed using these methods. The data set reported in detail here had 1870 values recorded as missing. These missing values were such that only five individuals had all attributes observed. One individual had all 12 attributes recorded as missing and was deleted from further analysis.

The mixture method of clustering requires the specification of the number of underlying number of clusters to be fitted to the model. Determination of the appropriate number of underlying clusters is still an unresolved problem, and there does not appear to be a universally superior method of determining the cluster number (see for example, Celeux and Soromenho (1996) and the references therein). In this paper, the problem of determining the group number is peripheral to the theory being presented, and we shall consider fitting two clusters to the model. This decision was based on having the clinical classification of the data into two groups, Stage 3 and Stage 4. We shall examine the extent to which the proposed techniques can rediscover these stages.

Hunt and Jorgensen (1999) report a complete case clustering of the 12 pre-trial covariates where individuals that had missing values in any of these covariates were omitted from further analysis, leaving 475 out of the original 506 individuals available. Hunt (1996) and Hunt and Jorgensen (1999) discuss a fitting strategy for incorporating local associations within the model. We report the results for the model where the three attributes WT, SBP and DBP are in one partition cell. This was the partitioning preferred by Hunt (1996).

We regard the data as a random sample from the distribution

$$f(\mathbf{x};\boldsymbol{\phi}) = \sum_{k=1}^{2} \pi_k \prod_{l=1}^{10} f_{kl}(\mathbf{x};\boldsymbol{\theta}_{kl}),$$

where the component distributions $f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl})$ are $N_3(\boldsymbol{\mu}_{kl},\Sigma_{kl})$ for the partition cell containing WT, SBP and DBP, $N(\mu_{kl},\sigma^2_{kl})$ for the remaining continuous attributes, and $D(\lambda_{kl1},\ldots,\lambda_{klm_l})$ for each of the four categorical attributes.

This model was fitted iteratively using the EM algorithm with an initial grouping based on the clinical classification. In the search for other maxima, the model was also fitted from a number of starting values generated by splitting the individuals into two groups both randomly and using various criteria. The first M step is then performed on the basis of these initial groupings. For discrete partition cells the initial estimates of the probabilities λ_{klm}, $m = 1,\ldots,M_l$ are calculated using the available data. For multivariate normal cells, the estimates of the means are calculated using the available data for that cell and in that group. The estimates for the variance covariances are calculated in this first M step by replacing the missing values in the cell by the group mean for that cell and then calculating the estimates using the 'filled in' data set. The convergence criterion used was to cease iterating when the difference in log-likelihoods at iteration t and iteration $t-10$ was 10^{-10}. Several local maxima were found, and the solution of the likelihood was taken to be the one corresponding to the largest of these. Each individual was assigned to the group to which it has highest estimated posterior probability of belonging.

Table 1
Agreements and differences between the clinical and model classifications

Clinical classification	Model classification		
	Class	Group 1	Group 2
	Stage 3	265	26
	Stage 4	41	173

It can be seen from Table 1 that the clinical classification and the 'statistical diagnosis' are different for 67 individuals. Examination of the posterior probabilities showed that 19 of these individuals are decisively assigned to a different group than the one corresponding to the clinical classification and nine have greater posterior probabilities lying between 0.5 and 0.6. Another comparison between the clinical classification and the model fit can be obtained by comparing the estimated parameters for the model with their counterparts using the clinical classification. Agreement was fairly close.

Hunt (1996) found in the complete case analysis that 40 of the 475 individuals were assigned to a different group than the one corresponding to the clinical classification. She found that survival status gave insight into the model classifications and the differences between the 'statistical diagnosis' and the clinical classifications. The model classification gave a better indication of prognosis with patients in Group 1 having a higher probability of being alive or dying from other causes, whereas patients in Group 2 had more chance of dying from prostatic cancer.

4. Discussion

When clustering real multivariate data sets having large numbers of attributes, it is rare that all variables are either categorical or continuous as some approaches based on finite mixture models require. The *Multimix* approach allows the clustering of mixed data containing both types of variables.

Missing values are also a problem in many classification studies. The lack of a formal model restricts the number of approaches that can cope with incomplete datasets. The finite mixture model leads itself well into coping with missing values. We have a well specified, yet flexible, model whose parameters can be estimated by maximum likelihood. As we have shown, the fitting method is able to be extended to cope with unintended missing data.

The approach implemented in this paper works extremely well for the mixed data set that had a very large amount of missing data. The model has performed well, detecting the structure known to exist in the data whilst simultaneously coping with an extreme amount of missing data. The parameter estimates for the clusters and the estimates of the missing attributes conditional on the group assignment are reasonable. However, as with all problems involving incomplete data, the mechanism that gives rise to the missing values does need careful investigation.

References

Afifi, A.A., Elashoff, R.M., 1966. Missing observations in multivariate statistics I: review of the literature J. Amer. Statist. Assoc. 61, 595–604.

Andrews, D.A., Herzberg, A.M., 1985. Data: A Collection of Problems from Many Fields for the Student and Research Worker. Springer, New York.

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

Beaton, A.E., 1964. The use of special matrix operators in statistical calculus. Educational Testing Service Research Bulletin, RB-64-51.

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

Celeux, G., Soromenho, G., 1996. An entropy criterion for assessing the number of clusters in a mixture model. J. Class. 13, 195–212.

Day, N.E., 1969. Estimating the components of a mixture of normal components. Biometrika 56, 464–474.

Dempster, A.P., 1969. Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading, MA.

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

Edwards, D., 1995. Introduction to Graphical Modelling. Springer, New York.

Everitt, B.S., 1984. A note on parameter estimation for Lazarsfeld's latent class model using the EM algorithm. Multivariate Behavioral Res. 19, 79–89.

Gordon, A.D., 1999. Classification. Chapman & Hall, CRC press, London.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27, 857–874.

Hartley, H.O., Hocking, R.R., 1971. The analysis of incomplete data. Biometrics 14, 174–194.

Hunt, L.A., 1996. Clustering using finite mixture models. Ph.D. Thesis, Department of Statistics, University of Waikato, New Zealand.

Hunt, L.A., Jorgensen, M.A., 1999. Mixture model clustering using the multimix program. Austral. and New Zealand J. Statist. 41, 153–171.

Jorgensen, M.A., Hunt, L.A., 1996. Mixture model clustering of data sets with categorical and continuous variables. In: Proceedings of the Conference on Information, Statistics and Induction in Science, Melbourne, 1996, pp. 375–384.

Lauritzen, S.L., Wermuth, N., 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann. Statist. 17, 31–57.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

Little, R.J.A., Schluchter, M.D., 1985. Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 72, 497–512.

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

McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Molenberghs, G., Goetghebeur, E.J.T., Lipsitz, S.R., Kenward, M.G., 1999. Nonrandom missingness in categorical data: strengths and limitations Amer. Statist. 53, 110–118.

Orchard, T., Woodbury, M.A., 1972. A missing information principle: theory and applications. Proceedings of the Sixth Berkeley Symposium, Vol. 1, pp. 697–715.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–593.

Schafer, J.L., 1997. Analysis of Incomplete Data. Chapman & Hall, London.

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