Mixture Model Clustering of Data Sets with Categorical and Continuous Variables

Murray A. Jorgensen maj@waikato.ac.nz
Lynette A. Hunt lah@waikato.ac.nz
Department of Statistics
University of Waikato ph: +64-7-856-2889
Hamilton, New Zealand fax: +64-7-838-4666

Abstract: *We introduce a class of multivariate mixture models that includes latent class models and mixtures of multivariate normal distributions as special cases. These models are fitted by maximum likelihood using the EM algorithm but the emphasis is less on parameter estimation than on the use of the estimated component distributions to cluster the data. To demonstrate the potential for application of these methods we use the program to fit several models to a large medical dataset.*

Keywords: Cluster analysis, EM algorithm, Latent class analysis, Local independence, Location model..

Area of Interest: Concept Formation and Classification.

1 Introduction

Two major difficulties frustrate the application of multivariate normal mixture models to clustering. Firstly, they are not easily adapted to cope with discrete data. This is unfortunate because most real clustering problems involve both continuous and discrete variables. Secondly, they lead to models with large numbers of parameters: for example if there are 8 variables we will need to estimate 36 parameters for even a common covariance matrix, many more if they must be estimated separately for each group.

Highly parameterized models can lead to difficulties in several ways. As discussed by McLachlan and Basford[13] (p. 11) the likelihood function of a mixture model can have singularities in a neighbourhood of which it is unbounded. Iterative methods for computing maximum likelihood estimates are drawn towards these singularities from many starting values if the model is highly parameterized. It is also common to find many local maxima in such models. Even if we find the largest of the local maxima we will often find the likelihood nearly constant in a low-dimensional set containing the local maximum.

When background information is available from subject-area theory or from previous statistical analysis of similar datasets it may well be possible to specify component distribution functions that are not highly parameterized and that are believed to represent the true shape of the components. We are concerned, however, with exploratory data analyses where very little may be known *a priori* about the structure of the data. What we need is a flexible, but not overly flexible, family of multivariate distributions that we can use as a 'default' for the component distributions in the absence of knowledge that would justify a more detailed specification. We draw our inspiration from Latent Class Analysis.

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 categorical variables indicating these states. Latent class models can be described as follows: we assume the population to be made up of K groups or sub-populations $G_1, \ldots, G_K$ in proportions $\pi_1, \ldots, \pi_K$. Let $\mathbf{x}$ be the vector of responses on the p variables that we observe on each observation, where the jth variable can take on levels numbered from 1 to M_j. If the ith observation $\mathbf{x}_i$ happens to come from G_k then its probability function is given by

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

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

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

so that the latent class model is a finite mixture model. The parameter vector $\boldsymbol{\phi}$ 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.

The original method of fitting these models, discussed at some length by Lazarsfeld and Henry [10] for the case of binary variables, was to attempt to solve the system of equations given by equating the predicted cell probabilities to the observed cell proportions. The solution of these equations can be difficult and Latent Class analysis became much easier to use when Goodman[4] introduced a new iterative algorithm for the maximum likelihood fitting of latent class models. It soon became clear that this algorithm was a special case of the very general EM algorithm discussed by Dempster, Laird and Rubin[3].

2 A general approach to multivariate mixture models

We will now sketch out a general class of multivariate mixture models for multivariate observations on both categorical and continuous variables. More specifically within this class we will describe the class of models that we actually use which generalises both latent class and multivariate normal 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

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

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

The distributions $f_k(x)$ 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 visualizing the data. Secondly the f_k must be restricted if we are to be able to have any hope of identifying the mixing proportions 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 $\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_p)'$ has been partitioned so that

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

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 *cell* of the partition, or simply 'the lth cell' if the partition being referred to is clear.

The form of local independence that we are assuming is that within each of the K subpopulations the variables in the cell $\breve{\mathbf{x}}_l$ are independent of the variables in $\breve{\mathbf{x}}_{l'}$ 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 $1, \ldots, M_j$ with probabilities $\lambda_{kl1}, \ldots, \lambda_{klM_j}$. We will denote this distribution by $D(\lambda_{kl1}, \ldots, \lambda_{klM_j})$. 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, x_j, and p_l continuous variables as elements. The discrete random variable takes values $1, \ldots, M_j$ with probabilities $\lambda_{kl1}, \ldots, \lambda_{klM_j}$. Conditional on the discrete variable taking value m the p_l continuous random variables have the multivariate normal distribution $N_{p_l}(\boldsymbol{\nu}_{mkl}, \Xi_{kl})$.

We can write the model for the ith observation as

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

where $\boldsymbol{\theta}_{kl}$ consists of the parameters of the distribution f_{kl} as described above. This model has been used for multivariate data with both categorical and continuous variables by Olkin and Tate [14], Krzanowski [7], and Little and Schluchter [12]. A referee has pointed out that Lawrence and Krzanowski [9] also consider the fitting of finite mixtures of location models. Strictly speaking the location model in full generality can have several categorical variables but for programming convenience we have reduced this to one. Location models are termed *homogeneous conditional Gaussian* by Lauritzen and Wermuth[8].

Note that in each of the K classes or subpopulations the vector random variable $\breve{\mathbf{x}}_l$ of the lth cell has the same type, either (a) or (b) or (c), but the parameters may vary from group to group. In fitting the model to a particular data set we have considerable discretion in how we form the $\breve{\mathbf{x}}_l$. 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 cell for the next series of fits, and so on.

Although we prefer to think of our models as mixture models it is interesting to note that they can be described in the language of graphical models used by Lauritzen and Wermuth[8] : 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 cell form a *clique* (maximal complete subgraph), all variables are connected to the latent variable, and variables in different cells are connected to each other only by a path through the latent variable.

3 Estimation and the MULTIMIX program.

As the model has been described, it is a mixture of K distributions, each of which can be seen to belong to the exponential family. It is therefore well suited for maximum likelihood estimation of its parameters by the EM algorithm of Dempster, Laird and Rubin [3], and the Fortran program MULTIMIX has been written by Lynette Hunt to do this. As is well known 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' is an extra variable giving the assignment of each observation to a class.

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. The 'complete data', then, consists of the $n \times p$ array of observed data $\{x_{ij}\}$ and the conceptual $n \times K$ array $\{z_{ik}\}$ of class membership indicators. The indicator vectors $\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_n$ are independently and identically distributed according to a multinomial distribution generated by one draw on a population made up of K categories in proportions $\pi_1, \ldots, \pi_K$.

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

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

where

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

Maximising the complete data log-likelihood $L_C(\boldsymbol{\phi})$ is equivalent to maximising $l_k(\boldsymbol{\theta}_k)$ separately for each cell. The local independence principles embodied in our models thus effectively reduce the dimensionality of the model-fitting, as well as improving the identifiability of the mixture components.

The 'missing data' formulation of the EM algorithm has made it possible to extend MULTIMIX to situations where the data are missing at random in the sense of Little and Rubin[11]. We add the genuinely missing components of the $\mathbf{x}_i$ to the $\mathbf{z}_i$ as data to be estimated at the E-step of the fitting algorithm. More strictly what are estimated at the E-step are the functions of the missing quantities as they appear in the sufficient statistics for the complete data log likelihood. Details are given by Hunt [6].

4 Does MULTIMIX give useful clusters? A medical example

There can never be one 'correct' method for performing a vaguely defined task like clustering. MULTIMIX clusters are based on maximum likelihood estimation of a parametric model, so one way to validate the program is to look at the performance of the program on data generated from the model. This was done repeatedly during the development of the program as a check on the code and, except where coding errors were indeed detected, the program performed well. Another method, using real data, is to withhold some variables from a cluster analysis and then examine whether the clusters found have any relationship with the excluded variables.

In this paper we consider the clustering of cases on the basis of pre-trial covariates alone for the Prostate Cancer clinical trial data of

Byar and Green [2] reproduced in Andrews and Herzberg[1], (pp. 261–274).

This data was obtained from a randomized clinical trial comparing four treatments for 506 patients with prostatic cancer grouped on clinical criteria into stages 3 and 4 of the disease. As reported by Byar and Green Stage 3 represents local extension of the disease without evidence of distant metastasis, while Stage 4 represents distant metastasis as evidenced by elevated acid phosphatase, x-ray evidence, or both. We will compare the clusters obtained by MULTIMIX with the clinical stages, and also consider the trial outcomes for patients in different clusters.

There are twelve pre-trial covariates (Table 1) measured on each patient, seven may be taken to be continuous, four to be discrete, and one variable (SG) is an index nearly all of whose values lie between 7 and 15, and which could be considered either discrete or continuous. We treat SG as a continuous variable. A preliminary inspection of the data showed that the size of the primary tumour (SZ) and serum prostatic acid phosphatase (AP) were both skewed variables. These variables have therefore been transformed, SZ under a square root transformation, and AP using a logarithmic transformation, to normalize their distributions. (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 fact several of the analyses to be described were also carried out using the version of the program which allows for missing observations. There was little variation from the results using only the complete observations.

Firstly we will consider two-group models: these have especial interest because of the clinical division into Stage 3 and Stage 4. Model-based classifications can be compared with this clinical classification.

We regard the data as a random sample from the distribution

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

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, which we will refer to as Model 1, 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, cell l; and $k = 1, 2$. We see that $f_{kl}(\breve{\mathbf{x}}_{il};\boldsymbol{\theta}_{kl})$ is $N(\mu_{kl}, \sigma^2_{kl})$ for each of the 8 continuous variables, and $D(\lambda_{kl1}, \ldots, \lambda_{klm_l})$ for each of the 4 categorical variables.

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. As the likelihood equation for mixture models usually has multiple roots, the EM algorithm should be applied from several starting values in order to search for local maxima. In order to search for other maxima, and to dispell any suspicion that the estimated parameters are close to the statistics for the clinical classification merely because these were used as starting values, the algorithm was run again 10 more times from initial parameter estimates taken from classifications generated by randomly splitting the patients into two groups. Three solutions of the likelihood equation were found for Model 1. From 10 starting values, 7 converged to a solution with a log-likelihood 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 log-likelihood of -11476.051, and one iteration

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

Table 1: Pretreatment covariates.

converged to a solution with a log-likelihood 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.

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 to the population G_k if $\tau_k(\mathbf{x}_i;\hat{\boldsymbol{\phi}}) \geq \tau_{k'}(\mathbf{x};\hat{\boldsymbol{\phi}})$ where $\tau_k(\mathbf{x}_i;\boldsymbol{\phi})$

$$\begin{aligned} &= \operatorname{pr}\left(i^{th}\text{observation} \in G_k|\mathbf{x}_i,\boldsymbol{\phi}\right) \\ &= \pi_k f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)/\{\sum_{k=1}^{2}\pi_k f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\}. \end{aligned}$$

On examination of the within group correlation structure (using the group assignment resulting from 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.

This correlation is incorporated into a new model in which variables SBP and DBP are grouped together in a cell, which we will refer to as Model 2. This model is a mixture of two component distributions, each of which is a product of 4 discrete distributions, 6 univariate normal distributions, and one bivariate normal distribution (for the blood pressures). The local independence condition 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 log-likelihood converged to -11268.723. Two other solutions (log-likelihoods -11275.551 and -11358.818) of the likelihood equation were also found when the EM algorithm was applied from a wide variety of starting values in the search of local maxima.

Using the cluster assignment from Model 2, the within group correlation structure is reexamined. There are small correlations between Wt and the blood pressures 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 involving some of these correlations.

The two group mixture model is fitted with the variables Wt, SBP and DBP grouped together in a cell (Model 3), and with the variables Wt and HG grouped in one cell, and SBP and DBP grouped together in another cell (Model 4).

Table 2 compares the classifications of the observations under the four models with the clinical classification.

The model classifications emerge as very similar to the clinical classification and seem to be little affected by the choice of model. In fact only a handful of observations change classification under the different models, these being observations 32, 58, 294 and 482. For observation 32 the estimated posterior probability of belonging to Group 1 (the less-seriously ill group) was 0.64, 0.58, 0.31 and 0.40 under Models 1 to 4 respectively. The corresponding probabilities for observations 58; 294; and 482 were 0.58, 0.62, 0.43 and 0.52; 0.49, 0.49, 0.51 and 0.45; and 0.49, 0.52, 0.45 and 0.42. None of these observations is decisively classified by any of the models, so from a clustering viewpoint the groups formed in this example are remarkably stable under these changes to the model.

In the same table we also indicate the improvement in fit gained by adding covariances to Model 1 by twice the log-likelihood ratio. Compared with Model 1, Model 2 has 2 extra parameters - one covariance between blood pressures for each of two clusters - and twice the difference in log-likelihoods is 235.1, clearly a significant improvement. Model 3 adds 4 extra parameters to Model 2 for a $-2\log\lambda$ gain of 28.0. Model 4 adds covariances between Wt and HG to Model 2 gaining 29.3 in $-2\log\lambda$ at a cost of 2 parameters. Both Model 3 and Model 4 offer significantly better fitting models than the fully locally independent model for a modest number of extra parameters. We do not recommend going too far in the direction of adding covariance parameters for fear of upsetting the stability of the model classifications. We have tended to prefer Model 3 on physical grounds because we would expect correlations between patient weight and the two blood pressures. We will remain with the covariance structure of Model 3 as we investigate adding more groups to the model.

4.1 Choosing the number of groups.

In many situations in practice, there is no *a priori* knowledge of the number K of component groups in the data. 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. The main problem is the lack of identifiability of the parameters even when the class of mixtures is identifiable. See for example, Hartigan[5], Titterington, Smith and Makov[16], and Quinn, McLachlan and Hjort[15].

We will use the likelihood ratio test merely a guide to the possible number of underlying groups. 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. The likelihood ratio test of $H_0 : K = 1$

	Stage 3		Stage 4		Number of	
Model	*Group 1*	*Group 2*	*Group 1*	*Group 2*	Parameters	2 logLR
1	252	21	20	182	55	0.0
2	252	21	21	181	57	235.1
3	252	21	18	184	61	263.0
4	252	21	19	183	59	264.3

Table 2: Comparison of 2-group models

versus $H_a : K = 2$ suggests the rejection of the null hypothesis of a single population ($-2 \log \lambda = 823.2$), twice the difference in the number of parameters being 60. The test statistics for $K = 2$ versus $K = 3$ and for $K = 3$ versus $K = 4$ are 188.3 and 175.8 respectively. As more groups were included in the model, there seemed to be an increasing tendency to converge to a suboptimal local maximum. This was not unexpected, since each additional cluster requires an additional set of 30 parameters to be estimated. We are confident that the best endpoint was reached for the 2 cluster solution, fairly sure for the 3 group solution, but are not at all confident for the 4 cluster solution. Although likelihood singularities are possible with these models, we encountered no instances where the algorithm failed to converge in the sense of our criterion. For reasons of time it was not practical to investigate 5 cluster models as the number of number of possible model varients coupled with increased sensitivity to starting values would make this a lengthy task.

On examination of the posterior probabilities for the groups fitted, we find (Table 3) 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 ($\hat{\tau}_{ij} \geq 0.95$).

The two cluster model does give groups with better separation. In this analysis, the two clusters found largely agree with the clinical classification of Stage 3 and Stage 4. When a 3 cluster model was fitted most of the Stage 4 patients were assigned to a single cluster, with the bulk of the Stage 3 patients being divided between the two other clusters.

	No. of Groups		
$\hat{\tau}_{ij}$	2	3	4
.25-.80	33	97	140
.80-.95	44	100	134
.95-.99	46	63	84
.99-1.0	352	215	117

Table 3: Posterior probabilities for 2-4 groups

4.2 Clusters and outcomes

We may gain additional insight into the composition of the groups from examining the cause of death. We will do this only informally as a detailed analysis of the data will take us too far from our main purpose. In particular we will neglect the treatment effects. Following [2], 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).

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

Model 3 uses a partitioning of the variables in which Wt, SBP, and DBP share a cell but

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

Table 4: Survival Status for Model 3 classifications

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

Table 5: Survival Status for a 3-group model

all other variables are locally independent. In Table 5 we consider the 3 group model with the same partitioning. Group 2 for this model corresponds roughly to the clinical classification of Stage 4 (Group 2 in the 2 cluster solution). The patients in Group 1 have a high probability of being alive at the end of the trial whereas the Group 3 patients have a high probability of death from cardiovascular causes, and similar moderate probabilities of death from other causes and alive at the end of the trial.

The post-treatment 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 occured, we will categorize this variable to survival time greater than 48 months, and survival time less than or equal to 48 months. With the Group 1 patients, 47% survive for greater than 48 months, whereas only 26.6% 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

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

Table 6: Survival time for the observations classified by Model 3 to a different group than the clinical classification

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

Table 7: Survival status for the observations classified by Model 3 to a different group than the clinical classification

conflict with the clinical classification, in Table 6 we tabulate survival time against the model classification for these patients alone.

In Table 7 survival status information is summarised for the same patients.

Tables 6 & 7 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.

5 Scope of the method

For fully categorical datasets the method of Latent Class Analysis has become a popular method of discovering underlying cluster structure. The class of models that we have introduced and utilized in this paper forms a natural extension to this class to datasets containing both categorical and continuous variables. Like Latent Class models, our 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 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. This may be particularly attractive in situations where data is missing, because the program MULTIMIX has been written to cope with data missing at random whereas missing data often presents problems for deterministic clustering algorithms.

The choice of discrete and multivariate normal 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 was known about the nature of the distributions in subpopulations other types of distributions could be used in place of these.

Either taken as we have presented them, or modified to incorporate subject-area knowledge of distributions and parameters we believe that multivariate finite mixture models will prove an invaluable tool in exploring large complex datasets.

References

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

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

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

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

[5] J.A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In L. M. Le Cam and R. A. Olshen, editors, *Berkeley Conference in Honour of Jerzy Neyman and Jack Kiefer*, volume II, pages 807–810, Monterey, 1985. Wadsworth.

[6] L. A. Hunt. *Clustering using finite mixture models.* PhD thesis, University of Waikato, Hamilton, New Zealand, 1996.

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

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

[9] C. J. Lawrence and W. J. Krzanowski. Mixture separation for mixed-mode data. *Statistics and Computing*, 6:85–92, 1996.

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

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

[12] R. J. A. Little and M. D. Schluchter. Maximum likelihood estimation for mixed continuous and categorical data

with missing values. *Biometrika*, 72:497–512, 1985.

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

[14] I. Olkin and R. F. Tate. Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.*, 32:448–465, 1961.

[15] G.J. Quinn, B.G., McLachlan and N.L. Hjort. A note on the aitken–rubin approach to hypothesis testing in mixture models. *J. R. Statist. Soc. B*, pages 311–314, 1987.

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