Mixture Model Clustering with the *Multimix* Program

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

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

Abstract

Hunt [1996] has implemented the finite mixture model approach to clustering in a program called *Multimix*. The program is designed to cluster multivariate data with categorical and continuous variables and possibly containing missing values. In this presentation we describe the approach taken to the design of *Multimix* and how some of the statistical problems were dealt with. As examples of the use of the program we cluster a large medical dataset and a version of Fisher's Iris data in which a third of the values are randomly made 'missing'.

1 INTRODUCTION

The *Multimix* computer program, written in Fortran by Lynette Hunt, fits a mixture of distributions to multivariate data where the variables may be either continuous or categorical. The model fitted simultaneously generalises the Latent Class model and the mixture of multivariate normals model. Like either of these models *Multimix* can be used to form clusters by the Bayes allocation rule. This is the intended use of the program, although the parameter estimates can be used to give a succinct description of the clusters.

The program is designed to encourage the use of parsimonious models that explain the associations and covariances by the cluster structure; it favours mixture component models with independent variables. In fact the user specifies a partition of the variables into groups such that variables in different groups are independent in the component models.

Use of the *EM* algorithm, with its view of the observed data as being notionally augmented by missing information to form the 'complete data', gives a broad framework for estimation which is able to handle two types of missing information: unknown cluster assignment and missing data. Using the methodology of Little and Rubin [1987] in this way *Multimix* is able to handle missing data in a less *ad hoc* way than many clustering algorithms. The program runs in acceptable time with large data matrices (say hundreds of observations on tens of variables). Use of the missing-data facility increases execution time somewhat.

2 STRUCTURE OF THE MODELS FITTED BY *Multimix*

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

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

where $\boldsymbol{\theta}_{kl}$ consists of the parameters of the distribution f_{kl} and the π_k are the mixing proportions. This formulation includes the motivating examples of Latent Class analysis [Aitkin et al., 1981] and mixtures of multivariate normals [McLachlan and Basford, 1988] When a subvector contains only a single variable, that variable is independent of all other variables within each subpopulation. It is convenient to assume forms for the f_{kl}, and hence for the f_k, that belong to the exponential family. The model is then well suited for maximum likelihood estimation of its parameters by the *EM* algorithm of Dempster et al. [1977].

2.1 LOCAL INDEPENDENCE

A simple case of the model class (2.1) occurs when each $\breve{\mathbf{x}}_l$ consists of a single variable. In this case true local independence holds and all variables are independent of each other within each subpopulation, even though they may be strongly related within the population as a whole. A succinct description of the data may then be given by reporting the proportions within each component and individual distribution summary statistics for each variable within each component. When all variables are discrete the model is then known as the *Latent Class* Model.

A natural way to generalize the Latent Class model to multivariate data involving variables of different kinds, discrete and continuous, is to consider local independence models with a range of different univariate distributions allowed for each variable. This kind of model often leads to fairly good cluster assignments even when it does not fit very well and the number of parameters that need to be estimated is fairly small. However it is easy to manufacture artificial examples where poor results are obtained, such as elliptical clusters in different orientations.

2.2 UNSTRUCTURED MODELS

Another extreme is where we take $L = 1$ and make no assumptions about independence within clusters. This is the case when we seek to fit a mixture of p-variate normals to the data, estimating means and covariance matrices separately for each component.

These models may be worth considering when p is quite small, but not otherwise. Highly parameterized models are very difficult to estimate unless very large amounts of data are available. Effective failure of identifiability may occur in which quite different parameter values give nearly equal likelihood values.

Another problem is that for variables of different types no explicit and tractable multivariate distributions may be known. This is the case when within-cluster associations need to be modeled for several categorical and several continuous variables.

2.3 MODELS SUPPORTED IN *Multimix*

Only some distribution types are currently available in *Multimix* although it is hoped to progressively extend these. Currently available are

1. arbitrary discrete univariate distributions
2. multivariate normal distributions
3. location model distributions.

The location model is a joint distribution of a single categorical distribution with several continuous variables. Conditional on each value of the categorical variable, the continuous variables follow a multivariate normal distribution. The mean vector of the multivariate normal may depend on the categorical variable, but the covariance matrix is the same whatever the value of the categorical variable. See Krzanowski [1983] for details.

The user of *Multimix* must partition the variables into groups or 'subvectors' in such a manner that each subvector corresponds to an available model. For example in a data set with 2 categorical variables and 15 continuous variables we might divide the variables into

- a group containing a single categorical variable
- a group containing a categorical variable and three continuous variables
- a group containing four continuous variables
- eight groups each containing a single continuous variable

these groups would be associated with a discrete distribution, a location model, a multivariate normal distribution, and eight univariate normal distributions and then the corresponding *Multimix* model would be a finite mixture of distributions, all having a joint density of the same form: a product of the 11 densities given above.

The model fitting strategy for fitting a mixture that we employ is to first fit the model with full local independence (in which each subvector of variables is a singleton). The *EM* algorithm which is used to calculate the maximum likelihood estimates of the parameters also produces for each observation the estimated probabilities $\hat{z}_{ij}$ that the observation i belongs to component j. We may 'sharpen' the resultant fuzzy classification by allocating each observation to the component that it has the highest probability of belonging to. The clusters so constructed can be examined for within-cluster correlations or associations. The model may then be modified by coarsening the partition of the variables so that variables with within-cluster associations are grouped in the same partition. One difficulty that may arise is the discovery of within-cluster associations involving more than one categorical variable. These may be handled by replacing those categorical variables by a single new categorical variable indexing the cells of the multi-way table that they define, possibly with some appropriate pooling of cells to reduce the number of values of the new variable.

The *EM* algorithm for fitting finite mixtures [McLachlan and Basford, 1988] treats the assignments z_{ij} of observations to clusters as it they were missing data. The adoption of the framework for *Multimix* means that it has been possible to extend the approach to cope with missing values in the data as well as missing cluster assignments. In fact having two categories of missing data complicates the situation sufficiently that the version of *Multimix* that handles missing data is about twice the length of the original *Multimix*. The algorithm used is an extension by Hunt [1996] of that of Little and Schluchter [1985] to finite mixtures of distributions. The *E*-step of the *EM* algorithm used incorporates a very efficient 'sweeping' on augmented covariance matrices to estimate the missing portion of the complete-data sufficient statistics. It is useful to have both versions of the program available, because the speed of the simpler program on data without missing values is greater than that of the missing-data version, and the parameter estimates found using the complete cases and the simpler program usually make good starting values for the larger program.

3 SOME EXAMPLES OF CLUSTERING WITH *Multimix*

3.1 BYAR PROSTATE CANCER DATA

The ultimate test of any clustering methodology is whether the clusters that result have any value for the user. To examine whether the clusters formed by *Multimix* have any usefulness we have clustered Byar's Prostate Cancer data [Andrews and Herzberg, 1985, pp. 5-8] into 2 groups using 12 pre-treatment covariates; 8 continuous variables and 4 categorical variables with between 2 and 7 levels. This is a useful data set to test clustering programs on, because the patients are classified into Stage 3 and Stage 4 (more severe) of the disease, and post-trial information on the survival status of the patients is available. The Stage 4 patients had some signs of the cancer spreading to other parts of the body. The stages were not used in the clustering, nor was the information on post-trial status (alive/dead/cause of death).

We report only an outline of our analysis of this data set, more details are given in Hunt [1996] and Hunt and Jorgensen [1999]. We use only the 475 out of 506 patients with complete pre-trial information. The initial model fitted was a 2-component model with complete local independence.

The initial clusters found had a strong relationship to the clinical stages:

	Cluster 1	Cluster 2
Stage 3	252	21
Stage 4	20	182

Inspection of the clusters showed three continuous variables that appeared to be associated within clusters. These were Systolic Blood Pressure, Diastolic Blood Pressure and Body Weight Index (Weight corrected for height). The physical plausibility of correlations between these variables within clusters gave additional reason to modify the model partition of variables bringing these three variables together in a subvector. Thus six additional covariance parameters are now estimated, three for each cluster.

The modification of the model model makes little difference to the clusters, in fact only four patients change cluster. The connection with the clinical stages tabulated above changes to

	Cluster 1	Cluster 2
Stage 3	252	21
Stage 4	18	184

Each patient in the data set has a survival status recorded that was not used in the clustering. A useful grouping of the status values is into four categories: alive(0), dead from prostatic cancer(1), dead from cardiovascular causes(2), dead from other causes(3). The following table shows that the clusters found have some prognostic value.

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

We can go on to investigate the clusterings generated by fitting models with more components. In the case of a three component model based on the same partitioning of the variables clusters are generated with the following relationship to the outcomes:

	Survival Status			
Cluster	0	1	2	3
A	56	18	31	21
B	43	12	63	40
C	38	91	44	18

To a good approximation Cluster 1 from the two component model splits into Cluster A and Cluster B, and Cluster C is more or less the same as Cluster 2. We can describe Cluster A as the more healthy patients; Cluster B patients are less healthy, but with health problems other than prostate cancer dominating; Cluster C patients are the main group at risk from prostate cancer.

Hunt [1996] goes on to investigate 4- and 5-component mixture models for the prostate cancer data. Some difficulty was experienced in fitting 5-component models as the *EM* algorithm took a long time to converge and many local likelihood maxima where encountered.

It is clear that *Multimix* is discovering structure related to the prognosis of the patients even though the Survival Status information is not used by the program.

3.2 FISHER'S IRISES WITH MISSING VALUES

A less 'real' but more familiar example is now derived from the Fisher Iris data by randomly making values missing with probability 1/3. The results in a data set that would be challenging to most clustering algorithms, but which is clustered easily by *Multimix*.

In the Iris data measurements of four variables Sepal length, Sepal width, Petal length, Petal width are available for 150 irises, 50 each from each of three species *Setosa*, *Versicolor*, and *Virginica*. As may be noted by graphical exploration the *Setosa* species is relatively well-separated from the other two. In a first exercise we take the 100 observations for *Setosa* and *Versicolor* and make the data values missing with probability 1/3. This is quite an extreme amount of missing data, in fact one *Setosa* observation finishes up with all four variables missing!

We consider two models: the local independence model, which in this case is a mixture of two 4-variate normal distributions both having diagonal covariance matrices; and the unstructured model which is a mixture of two general 4-variate normal distributions. The 100 observations are assigned by *Multimix* to clusters that are related to the species as follows:

	Cluster 1	Cluster 2
Setosa	48	2
Versicolor	1	49

Actually the assignment is the same for all observations under the two models. For example the totally unobserved *Setosa* iris is assigned to *Versicolor* by both models as the estimated proportion of *Versicolor* is 0.5011, and 0.5064 under the local independence and unstructured models respectively.

The assignment probabilities $\hat{z}_{ij}$ are somewhat closer to 0 or 1 under the unstructured model as it is able to fit the data better.

Separating *Versicolor* and *Virginica* is a harder task, even when all data values are present, but we repeat the exercise above with this pair. With the local independence model the clusters found relate to the original species as follows:

	Cluster 1	Cluster 2
Versicolor	37	13
Virginica	6	44

and with the unstructured model the corresponding table is

	Cluster 1	Cluster 2
Versicolor	36	14
Virginica	6	44

Although the results for the two models look similar, in fact 17 observations change their assignment to clusters between the two groups.

4 COMPARISON WITH RELATED SOFTWARE AND FUTURE DIRECTIONS

4.1 *AutoClass*

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

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

In fact the first is the more crucial difference, because the *EM* algorithm at the basis of both programs accommodates both ML and MAP estimation. *AutoClass* compares different models by calculating an approximation to the marginal density of the observed data after the model parameters have been integrated out. In usual *EM* language the approximation used is analogous to taking observed data likelihood to be proportional to complete data likelihood with the constant of proportionality to be evaluated at the maximum likelihood estimates.

The models currently available in *AutoClass* for attributes within a component are as follows. Categorical attributes are modelled by general discrete distributions (multi-category Bernoulli) as in *Multimix*. Continuous attributes may be taken to have uniform or normal distributions, possibly after transformation. Poisson distributions are available for count attributes. Cheeseman and Stutz [1996] report that von Mises-Fisher distributions for circular and spherical attributes are under development. At present it appears that *AutoClass* does not offer facilities for modelling within cluster dependencies, that is, all models assume within-cluster independence of attributes. Missing values are treated as a special kind of value in some attribute models, but there has been no implementation of the Little and Rubin [1987] methodology for data missing at random.

4.2 *Snob*

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

In contrast to *Multimix*, where the user must specify the number of classes, *Snob* selects the number of classes automatically using the Minimum Message Length criterion. Thus the MML criterion is used for all aspects of model selection and parameter estimation in the *Snob* approach.

4.3 *Mclust*

Banfield and Raftery [1993] have developed the classification likelihood approach of Scott and Symons [1971] further to introduce a controlled amount of flexibility to criterion-based cluster analysis for continuous data. Wallace and Dowe [1998] point out that in the case of a substantially overlapping pair of normal distributions having equal abundance and common σ this kind of estimation is likely to overestimate the difference in means and underestimate σ. This inconsistency in classification likelihood is also discussed by McLachlan and Basford [1988].

Banfield and Raftery characterize the dispersion matrices of multivariate normal clusters by their *orientation*, *size*, and *shape*. They mainly consider models where the shape is the same in each component of the mixture, but orientation and size are permitted to vary. They also consider an approach to robustifying cluster analysis by allowing a very dispersed 'noise' component in addition to the multivariate normal components.

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

5 THE PLACE OF *Multimix* IN MIXTURE MODELING

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

A natural further development for *Multimix* would be to introduce new types of attribute distribution such as the Poisson and circular von Mises distributions.

To the extent that robust estimation is appropriate for a particular dataset it seems that it would be better to add a very small proportion of a highly dispersed component to the mixture than to follow Banfield and Raftery [1993] in modifying the likelihood criterion to gain robustness.

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

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

References

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

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

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

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

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

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

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

L. A. Hunt and M. A. Jorgensen. Mixture model clustering using the *Multimix* program. *Australian and New Zealand Journal of Statistics*, 41(to appear), 1999.

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

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

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.

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

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

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

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

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