Statistics and Computing (1996) **6**, 85–92

Mixture separation for mixed-mode data

C. J. LAWRENCE and W.J. KRZANOWSKI

Department of Mathematical Statistics and Operational Research, University of Exeter, North Park Road, Exeter, EX4 4QE, UK

Received March 1995 and accepted June 1995

One possible approach to cluster analysis is the mixture maximum likelihood method, in which the data to be clustered are assumed to come from a finite mixture of populations. The method has been well developed, and much used, for the case of multivariate normal populations. Practical applications, however, often involve mixtures of categorical and continuous variables. Everitt (1988) and Everitt and Merette (1990) recently extended the normal model to deal with such data by incorporating the use of thresholds for the categorical variables. The computations involved in this model are so extensive, however, that it is only feasible for data containing very few categorical variables. In the present paper we consider an alternative model, known as the homogeneous Conditional Gaussian model in graphical modelling and as the location model in discriminant analysis. We extend this model to the finite mixture situation, obtain maximum likelihood estimates for the population parameters, and show that computation is feasible for an arbitrary number of variables. Some data sets are clustered by this method, and a small simulation study demonstrates characteristics of its performance.

Keywords: Cluster analysis, Conditional Gaussian distribution, EM algorithm, graphical modelling, location model, mixture maximum likelihood, simulation

1. Introduction

A common and very old problem in statistics is the separation of a heterogeneous population into more homogeneous subpopulations. A wide variety of approaches and techniques for tackling this problem now exists. Generic names for these techniques include classification, clustering and cluster analysis; for a concise account see, for example, Cormack (1971), Gordon (1981) or Everitt (1993).

We here focus on one specific approach, the mixture maximum likelihood method of cluster analysis originated by Day (1969) and Wolfe (1970), and described fully by McLachlan (1982) and McLachlan and Basford (1988). The mixture method is particularly suitable for clustering data sets that have too many individuals to be handled by pairwise distance algorithms; for a recent such application arising from analytical flow cytometry data, see Demers et al. (1992). In this method, the population of interest, π, is either known or assumed to consist of g different subpopulations $\pi_1, \ldots, \pi_g$, and the density of a p–dimensional observation $\boldsymbol{x}$ from π_i is assumed to be $f_i(\boldsymbol{x}; \boldsymbol{\theta}_i)$ for some unknown vector of parameters $\boldsymbol{\theta}_i (i = 1, \ldots, g)$. In this context the problem is to attempt a classification of a random sample of n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from π into the subpopulations to which they belong.

The mixture maximum likelihood method treats $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as a random sample of size n from a mixture of $\pi_1, \ldots, \pi_g$ in the proportions $\alpha_1, \ldots, \alpha_g (\sum_i \alpha_i = 1)$. The likelihood of this sample can be written

$$L_M(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g; \alpha_1, \ldots, \alpha_g) = \prod_{j=1}^{n} \left\{ \sum_{i=1}^{g} \alpha_i f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i) \right\}. \quad (1)$$

Assuming that each observation has an equal chance *a priori* of belonging to any of the g subpopulations, the posterior probability that $\boldsymbol{x}_j$ belongs to π_i can then be written

$$\tau_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i; \alpha_i) = \Pr\{x_j | \boldsymbol{\theta}_i; \alpha_i\} = \frac{\alpha_i f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i)}{\sum_{t=1}^{g} \alpha_t f_t(\boldsymbol{x}_j; \boldsymbol{\theta}_t)}. \quad (2)$$

Maximizing (1) with respect to the unknown parameters yields the maximum likelihood estimates $\hat{\boldsymbol{\theta}}_i, \hat{\alpha}_i$ for $i = 1, \ldots, g$. Setting $\hat{\tau}_{ij} = \tau_i(\boldsymbol{x}_j; \hat{\boldsymbol{\theta}}_i; \hat{\alpha}_i)$, the likelihood equations are given (McLachlan and Basford, 1988, Equations 1.6.1 and 1.6.2) by

$$\hat{\alpha}_i = \sum_{j=1}^{n} \frac{\hat{\tau}_{ij}}{n} \tag{3}$$

and

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \frac{\hat{\tau}_{ij}\partial \ln f_i(\boldsymbol{x}_j; \hat{\boldsymbol{\theta}}_i)}{\partial \hat{\boldsymbol{\theta}}_i} = \mathbf{0} \tag{4}$$

Observation $\boldsymbol{x}_j$ is then assigned to subpopulation π_k if $\hat{\tau}_{kj} \geq \hat{\tau}_{ij}$ for $i = 1, \ldots, g$.

For continuous data, normality of populations is a reasonable assumption to make in practice. The most general model is one in which the means and dispersion matrices are all allowed to differ in the subpopulations. However, allowing dispersion matrices to differ between subpopulations does have some theoretical drawbacks (such as singularities in the likelihood surface; see McLachlan and Basford, 1988, p. 38) as well as giving rise to too many unknown parameters when sample sizes are small. The most common approach in practice is therefore to assume that the mean vectors are different but the dispersion matrix is the same in all the subpopulations, i.e.

$$\boldsymbol{x}_j|\pi_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})(i = 1, \ldots, g). \tag{5}$$

In this case, the likelihood equations are given (McLachlan and Basford, 1988, pp. 38–42) by

$$\hat{\alpha}_i = \sum_{j=1}^{n} \frac{\hat{\tau}_{ij}}{n} \tag{6}$$

$$\hat{\boldsymbol{\mu}}_i = \sum_{j=1}^{n} \frac{\hat{\tau}_{ij}\boldsymbol{x}_j}{n\hat{\alpha}_i}, \tag{7}$$

and

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{g}\sum_{j=1}^{n} \frac{\hat{\tau}_{ij}(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)'}{n} \tag{8}$$

where

$$\tau_{ij} = \frac{\alpha_i \exp\left\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)\right\}}{\sum_{t=1}^{g} \alpha_t \exp\left\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_t)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_t)\right\}} \tag{9}$$

These equations can be solved iteratively by substituting initial estimates into the right-hand sides to produce new estimates on the left-hand sides, which are then substituted into the right-hand sides, and so on. These iterative estimates can be shown to be the same as those obtained using the EM algorithm (Dempster *et al.* 1977), so that convergence is guaranteed. However, this convergence may be to a local maximum, so several repeats of the process should be conducted from different starting values to ensure that a global maximum is attained. Also, convergence may be very slow; Day (1969) demonstrated that computing could be speeded up by a suitable reparametrization in the case $g = 2$, but there is no similar possibility in the general case.

The main practical drawback with the above scheme is that it is applicable only when the observed variables are continuous (as otherwise the normality assumption is not suitable), but in many applications some of the variables are either binary or categorical. There is thus need for a corresponding scheme with such mixed-mode data. Everitt (1988) proposed, and Everitt and Merette (1990) studied, an extension of the above model for such data. This extension assumes that an appropriate model for the data is given by a set of multivariate normal variables in which some of the continuous variables are only observed in categorical form because of thresholding. The thresholds that define categories are extra parameters of the model. The likelihood equations for mixture separation can be derived readily for this extended model, but they now contain multivariate normal integrals over as many variables as are thresholded with limits of integration depending on these thresholds. Thus each iteration of the process includes some additional multivariate quadrature, and as the limits of integration depend on unknown parameters this quadrature has to be fully repeated in each cycle. Consequently, this model is only a practical proposition if the number of categorical variables is relatively small.

In this paper, we therefore propose an alternative scheme for mixed-mode data. The model is one that has proved very useful for a number of years in discriminant analysis when the data contain both continuous and categorical variables (Krzanowski, 1993) but has come into prominence more recently because of its appearance in the graphical modelling of mixed-mode data (Whittaker, 1990). In Section 2 we outline the background, describe the model, and develop the mixture maximum likelihood theory. Some illustrative data sets are analysed in Section 3, and some characteristics of the method are investigated by means of the small simulation study described in Section 4. We conclude the paper with a discussion of the remaining problems.

2. Theory

In graphical modelling, the Conditional Gaussian distribution has been advocated as a suitable model for mixed-mode data (Whittaker, 1990; Edwards, 1990; Cox and Wermuth, 1992). This model has various equivalent parametrizations and expressions, but from our perspective the following is the most convenient. Suppose the data comprise p continuous variables $u_1, \ldots, u_p$ and q categorical variables $v_1, \ldots v_q$. Let the kth categorical variable have c_k categories ($k = 1, \ldots, q$). Then there are $m = \prod_{k=1}^{q} c_k$ distinct patterns of categorical variable 'values', and the set of categorical variables can thus be replaced by a single m–cell multinomial variable, w say. Any associations among the original categorical variables are converted into relationships among the resulting multinomial cell probabilities. Now consider a random sample of size n for which the original $p+q$ variables have been observed, and write it according to the cell of w that each individual occupies. Then the sample can be denoted by

$$\boldsymbol{x}_{11},\ldots,\boldsymbol{x}_{1n_1},\ldots,\boldsymbol{x}_{m1},\ldots,\boldsymbol{x}_{mn_m},$$

i.e. $\boldsymbol{x}_{sj}$ is the vector of continuous variable values for the jth out of the n_s individuals in cell s of the multinomial $w(s=1,\ldots,m;\sum_{s=1}^{m} n_s = n)$. The Conditional Gaussian model assumes that the distribution of the continuous variables depends upon the multinomial cell into which the corresponding categorical variables place an individual. Specifically, it is assumed that $\boldsymbol{x}_{sj} \sim N(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, and that the probability of observing an individual in cell s of the multinomial variable is $p_s(s=1,\ldots,m)$.

In the mixture separation case, it is again assumed that the above random sample has been drawn from a mixture of the subpopulations $\pi_1,\ldots,\pi_g$, so a modification of the model is needed to take account of this extra grouping structure. The most general modification is just to allow each unknown parameter to vary arbitrarily from subpopulation to subpopulation. Thus we now assume that $\boldsymbol{x}_{sj} \sim N(\boldsymbol{\mu}_{is}, \boldsymbol{\Sigma}_{is})$ in subpopulation π_i, and that the probability of observing an individual in cell s of the multinomial variable is p_{is} in $\pi_i(s=1,\ldots,m; i=1,\ldots,g)$. This was the model assumed by Krzanowski (1983) when deriving distances between populations of mixed-mode data. It is a perfectly general, and satisfactory, model at the theoretical level but again, as noted by Krzanowski (1983), it will cause estimation problems when applied in many practical situations because of the large number of parameters it contains. In order to make satisfactory progress, we need to restrict the number of parameters in some way. Consideration of similar models in other familiar statistical techniques (e.g. multivariate analysis of variance) suggests that the most fruitful modification is to constrain all the dispersion matrices to be equal, i.e. to set $\boldsymbol{\Sigma}_{is}$ equal to $\boldsymbol{\Sigma}$ for all i, s. The resulting model is known as the *homogeneous* Conditional Gaussian model in graphical modelling, but was originally introduced by Olkin and Tate (1961) as the 'location model'. It has proved a very successful model for discriminant analysis of mixed-mode data (see Krzanowski (1993) for a survey of these uses) and so will be considered now for the mixture separation problem.

Assuming, therefore, that $\boldsymbol{x}_{sj} \sim N(\boldsymbol{\mu}_{is}, \boldsymbol{\Sigma})$ and that $\Pr(w \in \text{cell}\, s) = p_{is}$ in $\pi_i(s=1,\ldots,m; i=1,\ldots,g)$, the joint probability of observing an individual in cell s and having associated continuous variable vector $\boldsymbol{x}_{sj}$ is given by

$$f_i(\boldsymbol{x}_{sj}, w_s; \phi) = \frac{p_{is}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_{sj}-\boldsymbol{\mu}_{is})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{sj}-\boldsymbol{\mu}_{is})\right\} \quad (10)$$

in subpopulation π_i, where ϕ denotes the whole collection of unknown parameters.

Now associate the g (unobserved) indicator variables $z_{1sj},\ldots,z_{gsj}$ with each $\boldsymbol{x}_{sj}$ where $z_{isj}=1$ if $\boldsymbol{x}_{sj}\in\pi_i$ and $z_{isj}=0$ if $\boldsymbol{x}_{sj}\notin\pi_i$. Then the log-likelihood of the mixture sample can be written

$$\begin{aligned} L_c(\phi) &= \sum_{i=1}^{g}\sum_{s=1}^{m}\sum_{j=1}^{n_s} z_{isj}\{\ln\alpha_i + \ln f_i(\boldsymbol{x}_{sj}, w_s; \phi)\} \\ &= \sum_{i=1}^{g}\sum_{s=1}^{m}\sum_{j=1}^{n_s} z_{isj}\{\ln\alpha_i + \ln p_{is} + \ln h(\boldsymbol{x}_{sj}; \boldsymbol{\mu}_{is}, \boldsymbol{\Sigma})\} \end{aligned} \quad (11)$$

where the z_{isj} are missing values and $h(\boldsymbol{x}_{sj}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the p.d.f. of a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random vector. Structurally this log-likelihood is the same as the one for mixture separation in the multivariate normal case (McLachlan and Basford, 1988 p. 37); it just has an extra summation (over cells, s) and some extra parameters (p_{is}). Indeed, the similarity of structure can be emphasised by viewing the log-likelihood (11) as a mixture of $g \times m$ normal populations having the hierarchical structure of g populations (with mixing proportions α_i) and m subpopulations within each of these populations (with mixing proportions p_{is}). The EM algorithm can be applied to this log-likelihood in exactly comparable form to that of McLachlan and Basford (1988); for current parameter values $\phi^{(k)}$ the expectation step yields $z_{isj}^{(k)} = \mathrm{E}(z_{isj}|\phi^{(k)}) = \tau_{isj}(\phi^{(k)})$, i.e. the value of τ_{isj} in (16) below evaluated at $\phi^{(k)}$, and the maximization step yields $\phi^{(k+1)}$ by maximizing $L_c(\phi^{(k)})$ of (11) with z_{isj} replaced by $z_{isj}^{(k)}$. This procedure yields the set of maximum likelihood equations:

$$\hat{\alpha}_i = \sum_{s=1}^{m}\sum_{j=1}^{n_s}\frac{\hat{\tau}_{isj}}{n} \quad (12)$$

$$\hat{p}_{is} = \sum_{j=1}^{n_s}\frac{\hat{\tau}_{isj}}{n\hat{\alpha}_i} \quad (13)$$

$$\hat{\boldsymbol{\mu}}_{is} = \sum_{j=1}^{n_s}\frac{\hat{\tau}_{isj}\boldsymbol{x}_{sj}}{n\hat{p}_{is}\hat{\alpha}_i}, \quad (14)$$

and

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{g}\sum_{s=1}^{m}\sum_{j=1}^{n_s}\frac{\hat{\tau}_{isj}(\boldsymbol{x}_{sj}-\hat{\boldsymbol{\mu}}_{is})(\boldsymbol{x}_{sj}-\hat{\boldsymbol{\mu}}_{is})'}{n} \quad (15)$$

where

$$\tau_{isj} = \frac{\alpha_i p_{is}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_{sj}-\boldsymbol{\mu}_{is})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{sj}-\boldsymbol{\mu}_{is})\right\}}{\sum_{t=1}^{g}\alpha_t p_{ts}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_{sj}-\boldsymbol{\mu}_{t})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{sj}-\boldsymbol{\mu}_{ts})\right\}} \quad (16)$$

A full specification of the process requires a method for determining starting values for the iterations. The simplest starting values are obtained by choosing a random set of $\tau_{isj}^{(0)}$ satisfying $\Sigma_i \tau_{isj}^{(0)} = 1$ for all s and j. (This can be done easily by assigning a random number between 0 and 1 for each $\tau_{isj}^{(0)}$ and then normalizing them to satisfy the constraint.) An alternative possibility is to partition the sample randomly into g groups and to take starting values

$$\alpha_i^{(0)} = \frac{1}{g}, \tag{17}$$

$$p_{is}^{(0)} = \frac{1}{m}, \tag{18}$$

$$\boldsymbol{\mu}_{is}^{(0)} = \sum_{j=1}^{n_s} \frac{\mathbf{x}_{isj}}{n_{is}}, \tag{19}$$

and

$$\boldsymbol{\Sigma}^{(0)} = \sum_{i=1}^{g} \sum_{s=1}^{m} \sum_{j=1}^{n_s} \frac{\left(\mathbf{x}_{isj} - \boldsymbol{\mu}_{is}^{(0)}\right)\left(\mathbf{x}_{isj} - \boldsymbol{\mu}_{is}^{(0)}\right)'}{n}, \tag{20}$$

where $\mathbf{x}_{isj}$ denotes those observations in cell s that have been allocated to group i, and n_{is} is the number of such observations.

Iteration proceeds until successive sets of τ_{isj} values are equal to within preset tolerance limits for all i, s, j. Since τ_{isj} again represents the posterior probability that $\mathbf{x}_{sj}$ belongs to π_i, $\mathbf{x}_{sj}$ is allocated to π_k if $\hat{\tau}_{ksj} \geq \hat{\tau}_{isj} (i = 1, \ldots, g)$. Since the method derives from the EM algorithm, convergence is technically guaranteed from any starting values. The problem in practice is that the log-likelihood surface frequently turns out to be relatively flat but with many local maxima. It is therefore recommended that solutions be obtained for a number of different random initial settings, and the solution corresponding to the maximum L taken in order to maximize the chances of reaching a global optimum.

3. Examples

In developing discrimination methods for mixed-mode data based on the location model, Krzanowski (1975) analysed several data sets each of which comprised two *a priori* groups of individuals. To illustrate the foregoing theory we treat some of these data sets as if they were single heterogeneous collections of individuals, separate each of them into two subgroups using the method in Section 2, and compare the derived partitions with the original groupings.

Table 1. *Numbers of individuals and variables in each of the data sets*

Data set	n	n_1	n_2	q	p
1	40	20	20	3	7
2	93	63	30	3	4
3	62	38	24	4	8
4	186	99	87	3	6

Specifically, we consider data sets 1–4 in Krzanowski (1975) as these four sets all contained both binary and continuous variables. Table 1 gives the total number of individuals (n), the numbers in each of the *a priori* groups (n_1, n_2), the number of binary (q) and the number of continuous (p) variables in each set; for further details of the background to each study and to the nature of the variables observed, see Krzanowski (1975).

Each data set was analysed twice, using each of the two starting procedures (start A: random τ's; start B: random assignment) outlined in Section 2. In each case, division into two subpopulations was sought (i.e. $g = 2$). Moreover, each analysis was repeated for 100 different random starts, the values at convergence were noted for each run, and the optimum solution (i.e. the maximum likelihood) over the 100 runs was taken. For each optimum solution, the classification of individuals into the two subgroups was compared with the *a priori* grouping of the individuals (matching groups so as to minimize discrepancies between the two partitions). In addition, a search was made through the 100 runs for each initialization method in order to identify the solution which gave the best match between *a priori* and recovered classification, and the log-likelihood of this solution was also noted. Finally, the log-likelihood of the *a priori* grouping was recorded for each data set. Table 2 lists all these various log-likelihoods, while Table 3 shows the corresponding classification comparisons. The latter are given in the

Table 2. *Log-likelihoods (L) under various conditions for the four data sets*

	Data set			
	1	2	3	4
max L (start A)	−1049.81	−1421.00	−1678.69	−5725.40
L (best allocation, start A)	−1049.94	−1425.79	−1687.25	−5725.40
max L (start B)	−1049.81	−1423.36	−1688.13	−5725.40
L (best allocation, start B)	−1081.48	−1428.31	−1766.39	−5774.10
L (*a priori* groups)	−1127.50	−1512.42	−1810.05	−5943.40

Table 3. *Classification matrices for each situation in Table 2*

	Data set							
	1		2		3		4	
max L, start A	14	10	32	10	22	14	72	51
	6	10	31	20	16	10	27	36
L (best allocation start A)	14	10	43	14	25	9	72	51
	6	10	20	16	13	15	27	36
max L (start B)	16	11	54	21	19	16	40	26
	4	9	9	9	19	8	59	61
L (best allocation, start B)	15	7	56	22	24	6	58	34
	5	13	7	8	14	18	41	53

form of 2 × 2 arrays

$$\begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$$

in which n_{ij} denotes the number of individuals allocated to group i by the method but originating from *a priori* group j for $i, j = 1, 2$.

Experience with these data sets showed that convergence of the log-likelihood maximization typically took between 20 and 40 iterations of the EM algorithm, and that there was quite a high level of consistency in the values at convergence of both log-likelihood and parameter estimates over the 100 runs for each initialization method. Data sets 1 and 4 yielded the same maxima with both initialization methods, but random τ's provided a slightly better solution than random allocations for the other two data sets. In all four cases, the log-likelihood for the *a priori* grouping was quite a long way from the maximum achieved, which indicates some conflict between model and data. Of course, these are four data sets that have arisen from real applications so there is the possibility of data recording errors, outliers or various other contaminations. Nevertheless, the consistency over replicates and initialization methods is encouraging, and the mechanics of the process stood up well to the different data sets.

Of particular interest is the classification of individuals, as this is often the major objective of the analysis. One measure of performance of the method is the confidence with which individuals are assigned to groups, and this can be assessed by considering the $\hat{\tau}_{isj}$. Assignment of $\boldsymbol{x}_{sj}$ to subpopulation π_i is made with confidence if $\hat{\tau}_{isj}$ is much greater than all other $\hat{\tau}_{ksj}$ (i.e. close to 1), while the assignment is equivocal if two or more $\hat{\tau}_{isj}$ are close in value. Adapting the measures in McLachlan and Basford (1988, p. 126) to the mixed-mode setting, therefore, we assess the overall level of confidence of assignment by

$$T = \sum_{s=1}^{m} \sum_{j=1}^{n_s} \frac{(\max_i \hat{\tau}_{isj})}{n}, \tag{21}$$

and the group-specific levels of confidence by

$$T_i = \sum_{s=1}^{m} \sum_{j=1}^{n_s} \frac{(\hat{z}_{isj}\, \hat{\tau}_{isj})}{n\hat{\alpha}_i} \tag{22}$$

where $\hat{z}_{isj} = 1$ if $\hat{\tau}_{isj} \geq \hat{\tau}_{ksj} \forall k$ and 0 otherwise. The closer each of these quantities is to 1, the more 'definite' are the individual assignments to subpopulations. Table 4 presents these values for the maximum likelihood solutions with each starting configuration on each data set (i.e. corresponding to the first and third rows of Table 2). It is evident that the assignments are very clear cut for all four data sets.

Table 4. *Measures of confidence for assignment of individuals in the four data sets*

		Data set			
Start	Measure	1	2	3	4
A	T	0.9996	0.9965	0.9918	0.9983
	T_1	0.9993	0.9977	0.9872	0.9984
	T_2	0.9999	0.9915	0.9972	0.9980
B	T	0.9999	0.9975	0.9986	0.9987
	T_1	0.9999	0.9916	0.9986	0.9992
	T_2	0.9999	0.9989	0.9986	0.9982

Finally, turning to the actual comparison of groups to which individuals were assigned by the method and their original group membership, there was rather more variation between replications and between initialization methods than there had been for the log-likelihood values or for the parameter. In addition to the classification achieved by the random start which provided the maximum of the log-likelihood for each initialization method, Table 3 also gives the best classification (i.e. the one most closely matching the *a priori* grouping) over the 100 random starts for each initialization and the corresponding log-likelihoods appear in Table 2. While there is relatively little difference among the four log-likelihoods for each data set, there is considerable variation in the classifications of the individuals and hence in the 'errors' relative to the *a priori* groups. This feature reflects the volatility of the process. Of course, there is considerable overlap in the *a priori* groups (estimated error rates over a range of discriminant functions varying between 17% and 35% for the four data sets; see Krzanowski, 1975). Hence it would be unrealistic to expect very close agreement between known and recovered classification. Also, the 'true' grouping is never known in a practical application so the second and fourth rows of each table will never be computable–they merely represent the 'best' that could be achieved over these replicates. The classification errors that would be achieved are those computed from the first or third rows, and these all fall around the 30%–40% level (comparable to the discriminant function errors quoted earlier).

It might be noted that the quantities T, T_1 and T_2 defined above can also be used to estimate error rates when initial group membership is unknown. Specifically, $1 - T$ is an estimate of the overall error rate, while $1 - T_i$ is an estimate of the group-specific error rate for π_i. However, McLachlan and Basford (1988, pp. 126–32) point out that these estimates are optimistically biased (as is very evident from Table 4), and provide bootstrap schemes for correcting the bias. These schemes are very intensive computationally, however, so were not carried out here.

4. Monte Carlo study

A small Monte Carlo study was mounted in order to investigate some of the characteristics of the proposed methodology. Each replicate comprised 20 observations generated from each of two 4-variate normal populations. One population had mean vector (0,0,1,1), the other had mean vector (0,0,6,6), and both populations had common dispersion matrix

$$\begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 3 \end{pmatrix}.$$

The first two variates for each observation were dichotomized by thresholding: if x_{ij} is the generated value of continuous variable i in population j, then a binary value y_{ij} is produced by setting $y_{ij} = 0$ if $x_{ij} < w_{ij}$ and $y_{ij} = 1$ if $x_{ij} \geq w_{ij}$. Chosen thresholds were $w_{ij} = 0$ for all i, j. The resulting data set thus comprised 20 observations on $p = 2$ continuous and $q = 2$ binary variables from each of two populations. The mixture separation technique described above was then applied with $g = 2$ (i.e. partitioning into two groups); initialization of the iterative process was by means of random τ's (method A), and 50 random starts were included in each replication. The resulting clustering can be compared with the true grouping to determine the error rate of the procedure, and the estimated parameter values can be compared with the true values. The whole process was replicated 50 times to investigate variability under repeated sampling.

Figures 1 and 2 show histograms of the maximum log-likelihood values and the numbers of misclassified individuals over the 50 replications. The average maximum log-likelihood value was −212.30 with a standard error of 0.77, and the average number of misclassificiations was 12.56 with a standard error of 0.78. Thus an approximate 95% confidence interval for the error rate in a single application of the process lies between 27.5% and 35%. Note however that the histogram of maximum log-likelihood values exhibits a reasonably normal shape, but the one for the average number of misclassifications is much more uniformly spread. Also, for the underlying 4-variate normal population, the difference in (true) mean vectors is (0,0,5,5) and the inverse (true) dispersion matrix is

$$\frac{1}{9}\begin{pmatrix} 7 & -2 & -2 & -1 \\ -2 & 7 & -2 & -1 \\ -2 & -2 & 7 & -1 \\ -1 & -1 & -1 & 4 \end{pmatrix}.$$

This yields a (true) Mahalanobis D^2 between populations of 25, and hence the best possible discriminant procedure (using *identified* individuals and *true* parameter values) would still incur an error rate of $2\Phi(-(1/2)D) = 1.25\%$ (McLachlan, 1992, p. 17). In the present case we have

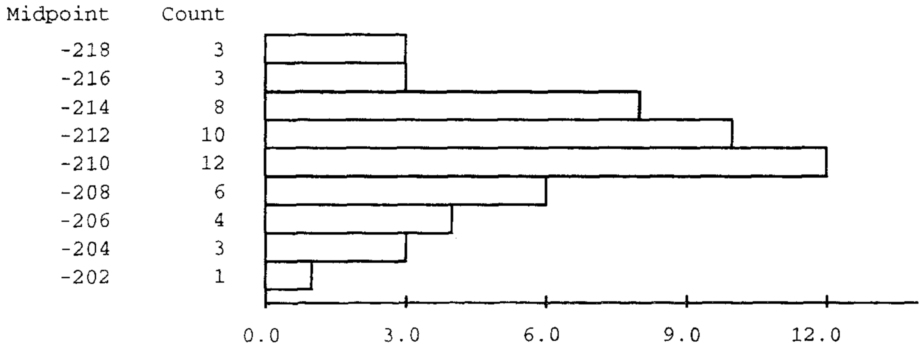

Fig. 1. *Histogram of the maximum log-likelihood values over 50 replications of the simulation experiment*

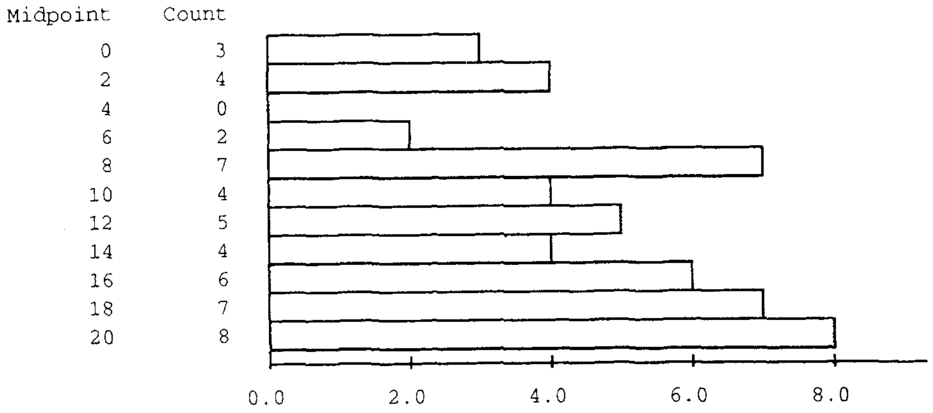

Fig. 2. *Histogram of the numbers of misclassified individuals over 50 replications of the simulation experiment*

unidentified training data and *samples* rather than populations, so the success rate of the classification is in fact very good.

A final point concerns the estimation of parameters. The two population mean vectors for the continuous variables were (1,1) and (6,6). Averages of their estimates over the 50 replications were (2.52, 2.53) and (4.47, 4.47) respectively, with standard errors around 0.1 for each element of each vector. There has thus been a shrinkage of about 1.5 towards the centre for both variates. This shrinkage is of course highly correlated with the error rate noted above, as misclassification of individuals will inevitably contract the difference between the groups. The dispersion matrix for the two continuous variates,

$$\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix},$$

had average estimate

$$\begin{pmatrix} 1.753 & 0.883 \\ 0.883 & 2.320 \end{pmatrix}$$

with standard errors

$$\begin{pmatrix} 0.185 & 0.173 \\ 0.173 & 0.119 \end{pmatrix}$$

over the 50 replicates.

5. Discussion

Experience with application of the methodology outlined here has been very encouraging to date, and relatively few problems have been encountered. Convergence of the EM algorithm has been very satisfactory for the scale of data sets analysed, with never more than about 60 iterations required, but the log-likelihood surface has typically contained many local maxima of fairly comparable heights. Thus, while there is a large measure of consistency in the actual maximum values over different random starts for a given data set, the parameter estimates and classifications of the individuals do show considerably more variation over these local maxima. It is thus important to ensure that the algorithm is run from plenty of different random starts; we used between 50 and 100 in all our analyses in order to be confident of reaching an optimum solution.

In all those analyses in which there existed an *a priori* grouping of individuals, matching recovered groups with original groups was not immediate and had to be done by trial and error. We always chose that matching which yielded fewest misclassifications, so there is liable to be some optimistic bias in our reporting. However, this matching is artificial in that in any practical application there is no *a priori* grouping, so the focus of interest is generally on the adequate *description* of the recovered groups from the estimated parameters of the model. Of course, the additional question of determining an optimum number of subgroups into which to divide the data raises exactly the same problems as in the continuous case (McLachlan and Basford, 1988), so no definitive method is available for the mixed-mode case either.

Computing times have been quick for all the analyses reported here, and there appears to be no reason why the method should not work equally well on arbitrarily large numbers of categorical and continuous variables. Of course, increasing the number of categorical variables automatically increases the number of multinomial cells in the Conditional Gaussian model, and hence increases considerably the number of parameters to be estimated. There must be a trade-off between number of parameters and sample sizes for viability of method, but we have not investigated this aspect yet. If samples are small then some constraints on the model parameters may be necessary (e.g. the sort of linear model structure that has proved valuable in discriminant analysis; see Krzanowski, 1993, for details). This is another question for future study.

We conclude with some general points of comparison between the Conditional Gaussian model used in this paper and the threshold model employed by Everitt (1988) and Everitt and Merette (1990). Both models assume normality of the continuous variables, and both either assume or imply common dispersion matrices for these variables across subpopulations. However, the threshold model also implies the existence of a continuous latent variable underlying the categories of each observed categorical variable, while the Conditional Gaussian model merely assumes a different conditional distribution of the continuous variables for each category combination of the categorical variables. Furthermore, the threshold model imposes several inherent orderings: of the categories in each categorical variable (induced by the cut-points of the latent variables) and of the conditional means of the continuous variables (due to the linear regression structure imposed by the normality assumption).

It is thus important to bear these constraints in mind when considering practical applications. In particular, the threshold model should be employed only in those situations where existence of an underlying latent continuum for the categorical variables is justifiable, and where the implied orderings make substantive sense. None of these constraints appears to be particularly restrictive in the special case of binary categorizations and, indeed, the threshold model has provided a satisfactory basis for multivariate probit analysis as well as for other techniques associated with dose-response relationships (see for example Ashford and Sowden, 1970). Greater care is needed for more general categorical variables, however. The Conditional Gaussian model, by contrast, appears to be defensible in quite general circumstances. Indeed, it will also provide satisfactory performance on data conforming to the threshold model (as demonstrated by the simulation results above).

References

Ashford, J. R. and Sowden, R. R. (1970) Multi-variate probit analysis. *Biometrics*, **26**, 535–46.

Cormack, R. M. (1971) A review of classification (with discussion). *Journal of the Royal Statistical Society*, Series A, **134**, 321–67.

Cox, D. R. and Wermuth, N. (1992) Response models for mixed binary and quantitative variables. *Biometrika*, **79**, 441–61.

Day, N. E. (1969) Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–74.

Demers, S., Kim, J., Legendre, P. and Legendre, L. (1992) Analyzing multivariate flow cytometric data in aquatic sciences. *Cytometry*, **13**, 291–8.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

Edwards, D. (1990) Hierarchical interaction models (with discussion). *Journal of the Royal Statistical Society*, Series B, **52**, 3–20.

Everitt, B. S. (1988) A finite mixture model for the clustering of mixed mode data. *Statistics and Probability Letters*, **6**, 305–9.

Everitt, B. S. (1993) *Cluster Analysis*, 3rd Edn. Edward Arnold, London.

Everitt, B. S. and Merette, C. (1990) The clustering of mixed-mode data: a comparison of possible approaches. *Journal of Applied Statistics*, **17**, 283–97.

Gordon, A. D. (1981) *Classification*. Chapman and Hall, London.

Krzanowski, W. J. (1975) Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, **70**, 782–90.

Krzanowski, W. J. (1983) Distance between populations using mixed continuous and categorical variables. *Biometrika*, **70**, 235–43.

Krzanowski, W. J. (1993) The location model for mixtures of categorical and continuous variables. *Journal of Classification*, **10**, 25–49.

McLachlan, G. J. (1982) The classification and mixture maximum likelihood approaches to cluster analysis. In P. R. Krishnaiah and L. N. Kanal (eds.), *Handbook of Statistics*, Vol. 2, pp. 199–208. North-Holland, Amsterdam.

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

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

Olkin, I. and Tate, R. F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**, 448–65 (correction **39**, 1358–9).

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Wolfe, J. H. (1970) Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, **5**, 329–50.