COENOSES 5 (1): 23-31, 1990

LIKELIHOOD APPROACH TO CLUSTERING USING UNCLASSIFIED INITIAL SAMPLE[1]

S. Ganesalingam, Department of Mathematics and Statistics, Massey University, Palmerston North, New Zealand

Keywords: Linear discriminant function, normal distribution mixture and classification clustering approaches, mixture and separate sampling, maximum likelihood estimates, simulation comparison

Abstract. Mixtures of distributions, in particular the normal, have been used extensively as models in a wide variety of important practical situations where the population of interest may be considered as consisting of two or more subpopulations mixed in varying proportions. The problem of decomposing such a mixture of distributions is of considerable interest and utility. Two commonly used clustering methods based on maximum likelihood are considered in the context of the classification problem where observations of unknown origin belong to one of the two possible populations. The basic assumptions and associated properties of the two methods are contrasted and illustrated by a series of simulations under two different sampling schemes, namely mixture sampling and separate sampling. A case study is presented to demonstrate the basic differences between these two methods.

Introduction

Large multivariate data sets can prove difficult to comprehend, and methods of summarizing and extracting relevant information are necessary. One approach is through clustering the data set into small groups. Clustering or classifying individuals into groups such that there is relative homogeneity within groups and heterogeneity between the groups is a problem which has been considered for many years. In general terms a clustering method attempts to group unclassified individuals into clusters so that members of the same cluster are, in some sense, similar to one another. The concept of clusters encompasses the duality of homogeneity within clusters and heterogeneity between clusters. The available methods of clustering may be categorized broadly as being hierarchical or non-hierarchical. The mixture and classification maximum likelihood approaches are examples of two non-hierarchical procedures in wide use today. The latter approach is also an example of a clustering technique which sets out to optimize a specified criterion for a given number of clusters. Another procedure of this type is the so-called k-means model of MacQueen (1966) and Hartigan (1975).

This paper examines the relative performance of these two non-hierarchical clustering methods based on maximum likelihood, namely mixture maximum likelihood approach (MA) and classification maximum likelihood approach (CA). We consider here the situation where the population of interst Π is known or assumed to consist of two normal subpopulations Π_1 and Π_2 with unknown means μ_1 and μ_2 and common covariance matrix Σ. Given a random sample of observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ from Π, the problem is to allocate each $\mathbf{y}_k$, $(k=1, 2, \ldots, n)$ to the subpopulations to which it belongs. We let $\gamma=(\gamma_1, \ldots, \gamma_n)$ be the set of identifying labels, where $\gamma_k=1$ or 0 according as $\mathbf{y}_k$ belongs to Π_1 or Π_2. This would be a classical discrimination problem if γ were known *a priori*. A discrimination procedure would be formed from the classified sample for the allocation of subsequent observations of unknown origin. In Section 2 we introduce briefly the two methods in a very general context where in addition to the unclassified data, the initial sample also contains some classified data

$$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i}, (i=1, 2); \; m=(m_1+m_2).$$

The case m = 0 is the extreme case facing statisticians, and an example is considered in the case study presented in Section 4. However, having some classified observations from each subpopulation (*i.e.*, $m_i>0$) facilitates the estimation process in a number of ways. For instance, they provide initial estimates of the subpopulation parameters to be used at any iterative process, and they may also prevent the occurrence of singularities in the likelihood which may otherwise occur when m = 0. Further, they allow checking of the subpopulation densities, say, for normality using methods such as proposed by Andrews (1973) and Gnanadesikan (1977). The case m > 0 is of interest when the aim is to estimate the mixing proportions where the *n* unclassified observations have been sampled from Π, but where the classified data have been obtained by separate sampling and so provide no information about

[1] Paper presented at the 2nd CETA International Workshop on Mathematical Community Ecology, Gorizia, Italy; 19-25 November 1988.

the mixing proportions.

2. The two approaches

We shall introduce here the two approaches MA and CA in the situation where the initial sample contains some classified observations $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, ... $\mathbf{x}_{im_i}$ from Π_i: $i=1, 2$, $(m=m_1+m_2)$ in addition to the n unclassified observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Although the simulation studies undertaken in this paper are for the case $m=0$, we consider the application of MA and CA for $m>0$, as the two methods are also applicable to this case.

2.1 The mixture approach (MA)

The likelihood of the $(m+n)$ observations is given by

$$L_M = \prod_{i=1}^{2} \pi_i^{m_i} \sum_{j=1}^{m_i} f_i(x_{ij}) \prod_{k=1}^{n} \{ \sum_{i=1}^{2} \pi_i f_i(y_k)\} \qquad (2.1.1.)$$

or

$$L_M = \prod_{i=1}^{2} \prod_{j=1}^{m_i} f_i(x_{ij}) \prod_{k=1}^{n} \{ \sum_{i=1}^{2} \pi_i f_i(y_k)\} \qquad (2.1.2)$$

according as the m_i classified observations are obtained by mixture or separate sampling, where f_i denotes the multivariate normal density with unknown means μ_i and covariance matrix Σ, and π_i $(i=1, 2)$ is the proportion in which y_k are obtained from the mixture Π. The values of the parameters μ_i, Σ, π_i are chosen to maximize (2.1.1) or (2.1.2) depending on the sampling schemes adopted to obtain the classified observations. The posterior probability that $\mathbf{y}_k$ belongs to Π_r is given by

$$\theta_r(y_k, \hat{\mu}_r, \hat{\Sigma}) = \hat{\pi}_r f_r(y_k, \hat{\mu}_r, \hat{\Sigma}) / \sum_{i=1}^{2} \hat{\pi}_i f_i(y_k, \hat{\mu}_i, \hat{\Sigma}); \qquad (2.1.3)$$

$r=1, 2;$

where $\hat{\mu}_i$, $\hat{\Sigma}$, $\hat{\pi}_i$ $(i=1, 2)$ are the mixture maximum likelihood estimates obtained using EM algorithm of Dempster et al (1977). For convenience we shall denote $\theta_i(y_k, \hat{\mu}_i, \hat{\Sigma}) = \hat{\theta}_{ik}$. The efficiency aspect of this mixture approach has been investigated by Ganesalingam and McLachlan (1978, 1979a) and O'Neill (1978).

In the case of two normal subpopulations with equal variance and covariance matrices, the allocation based on 2.1.3 simplifies to the use of a linear discriminant function

$$\mathbf{W} = \mathbf{a}'\mathbf{y} + \mathbf{b} \qquad (2.1.4)$$

where $\mathbf{a} = \Sigma^{-1}(\mu_2 - \mu_1)$ and

$$\mathbf{b} = \frac{1}{2}(\mu_1 + \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) + \log\left(\frac{\pi_2}{\pi_1}\right).$$

The population parameters μ_1, μ_2 and Σ and hence $\mathbf{a}$, and $\mathbf{b}$ are estimated by maximizing the appropriate L_M. We shall call the estimate of $\mathbf{a}$ and $\mathbf{b}$ thus obtained $\mathbf{a_M}$ and $\mathbf{b_M}$, and consequently (2.1.4) now reads

$$\mathbf{W_M} = \mathbf{a}'_{\mathbf{M}}\, \mathbf{y} + \mathbf{b_M} \qquad (2.1.5)$$

2.2 The classification maximum likelihood approach (CA)

With the classification maximum likelihood approach the labels γ_k are treated as unknown parameters. The likelihood of the $(m+n)$ observations is then given by

$$L_C = \prod_{i=1}^{2} \prod_{j=1}^{m_i} f_i(x_{ij}) \prod_{k=1}^{n} \prod_{i=1}^{2} \{f_i(y_k)\}^{(\gamma_k)_i} \qquad (2.2.1)$$

The subpopulation parameters and the identifying labels γ_k are chosen to maximize the likelihood given by (2.2.1). The maximization is over the set of values of the γ_k corresponding to all possible assignments of the $\mathbf{y}_k$ to the two subpopulations as well as over all admissible values of the parameters. In principle the maximization process can be carried out since it is simply a matter of computing the maximum value of the likelihood (2.2.1) over all possible partitions of the n unclassified observations to the two subpopulations. However, unless the total number of observations in the sample is quite small, the length of this search is prohibitive; see John (1970a, b).

It follows (John 1970a) that the estimate of γ_k, $\hat{\gamma}_k$, is 1 or 0 according to

$$W_C^o(\mathbf{y}_k) = \log(f_2(\mathbf{y}_k; \hat{\mu}_2, \hat{\Sigma}) / f_1(\mathbf{y}_k; \hat{\mu}_1, \hat{\Sigma})\} \qquad (2.2.1)$$

$$= \mathbf{a}'_C \mathbf{y}_k + \mathbf{b}_C^o$$

is less or greater than zero, where

$$\mathbf{a}_C = \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

and

$$\mathbf{b}_C^o = -\frac{1}{2}\mathbf{a}'(\hat{\mu}_1 + \hat{\mu}_2)$$

are the maximum likelihood estimates of a and $\mathbf{b}^\circ = \mathbf{b} - \log(\pi_2/\pi_1)$, and $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\Sigma}$ are the ordinary maximum likelihood estimates for the sample after the unclassified data are classified according to the γ_k. Hence the solution can be computed iteratively.

Unfortunately with CA, the γ_k increase in number with the number of unclassified observations, and under such conditions it is well known that the maximum likelihood estimates need not be consistent. It appears (Marriott 1975 and Bryant and Williamson 1978) that MA avoids the asymptotic biases associated with CA where, at each stage of the iterative process of computing the maximum likelihood estimates, each $\mathbf{y}_k$ is assigned outright to a particular subpopulation according to the estimate for γ_k (k = 1, ..., n). By contrast the iterative estimation process associated with MA does not insist on definite membership of $\mathbf{y}_k$ to any subpopulation. Rather at any stage it gives for each $\mathbf{y}_k$ an estimated probability of membership of each subpopulation. An excellent discussion of these two approaches for the case m = 0 was given recently by McLachlan (1980).

It can be seen that MA makes an additional assumption that $(\gamma_1)_i, \ldots, (\gamma_n)_i$ is an unobservable random process from a Bernoulli distribution with probability parameter π_i (i = 1, 2). If MA is adopted and this assumption is not satisfied, then the resulting estimates need not possess the desirable properties of maximum likelihood estimators under regularity conditions. However, we shall see that MA might still outperform CA as a clustering procedure in most circumstances.

2.3 Error rates

Let P_{ij} denote the probability that a randomly chosen member of Π_i (i = 1, 2) is misallocated by W_j for j = C, M. P_{iC} and P_{iM} are conditional on the initial sample estimates of a and b under the respective approaches. It follows for i = 1, 2 that

$$P_{iM} = \Phi\ [(-1)^{i+1}\,(a'_M\,\mu_i + b_M)\,/\,(a'_M \textstyle\sum a_M)^{1/2}\ ]; \qquad (2.3.1)$$

P_{iC} is defined by replacing $\mathbf{a}_M$ and $\mathbf{b}_M$ with $\mathbf{a}_C$, $\mathbf{b}_C$ respectively in (2.3.1).

For j = C or M the overall error rate associated with W_j is defined by

$$R_j = \pi_1\ P_{1j} + \pi_2 P_{2j} \qquad (2.3.2)$$

and $E(R_j)$ denotes the overall expected error rate obtained by averaging R_j over the sampling distribution of the appropriate estimates of **a** and **b**. Note that R_M and R_C refer to the overall error rates with respect to an observation subsequent to the initial sample. It would be extremely difficult to give a theoretical account of the conditional error rate associated with the application of W_j to one of the initial n unclassified observations owing to their correlation with $\mathbf{a}_j$ and $\mathbf{b}_j$, j = M, C.

The performances are also to be assessed on the basis of their overall apparent error rates R^*_M and R^*_C, obtained by applying W_M and W_C to the original sample on which they are based. R_j^* is given by

$$R_j^* = (n_{1j} + n_{2j})/n$$

where n_{ij} is the number of members from Π_i in the sample misallocated by W_j (i = 1, 2), that is, R_j^* denotes the proportion of members in the original sample misallocated by W_j.

3. Simulation experiments

In this section we wish to contrast the performance of MA with CA in the extreme case of m = 0, that is, when we have only n unclassified observations. For two subpopulations the label γ_k is either 1 or 0 according as y_k belongs to Π_1 or Π_2. It has been noted by Mariott (1971) and Scott and Symons (1971) that CA tends to produce groupings of roughly the same size. However, there appears to be very little information in the literature on the relative performance of MA and CA, particularly, as to which approach is preferable in general.

A series of simulation experiments were carried out where at each trial the two methods were applied to allocate the same sample simulated for both mixture and separate sampling schemes. In practice the type of sampling scheme may not be known to the statistician. It will therefore be of interest to see how satisfactory CA is under mixture sampling or MA under separate sampling. The relative performance of the two discriminant functions W_M and W_{C° as formed by MA and CA respectively will be assessed on the basis of their overall error rates R^*_M (R^*_C) and R_M (R_C) associated with the application of W_M (W_{C°) to the original sample and to a subsequent observation respectively. The convenient canonical form of $\mu_1 = -\mu_2 = (\frac{1}{2}\Delta,\ 0,\ \ldots,\ 0)'$ and $\Sigma = I$ was assumed without loss of generality, where Δ is the Mahalanobis distance between the two subpopulations Π_1 and Π_2 with means μ_i (i = 1, 2) and common covariance matrix Σ and is given by $\Delta = \{(\mu_1-\mu_2)'\ \Sigma^{-1}\ (\mu_1-\mu_2)\}^{1/2}$. Attention was focussed on dimensions p = 1, 2 and 4. Thirty simulated trials were performed over 42 different combinations of the parameters Δ, π_1, p, and n under both schemes of sampling (Δ = 1, 2, 3; π_1 = 0.25, 0.35, 0.5; n = 20,40). Under separate sampling π_i refers to the fixed proportion of the unclassified data coming from Π_i. That is, $n\pi_i$ observations were taken from Π_i on each trial (i = 1, 2), where π_i was chosen so that $n\pi_i$ was an integer.

For each combination of the parameters, the average overall error rates associated with W_M and W_{C° were obtained by computing the sample means R_j and R_j^* of the values of R_j and R_j^* respectively obtained on each trial (j = M, C). As noted earlier, MA produces multiple maxima for p > 3, and so in the case of p = 4 a systema-

Table 2. Error rates R_C and R_M for n = 40, p = 2.

		Sampling Scheme					
		Mixture			Separate		
			Δ			Δ	
		1	2	3	1	2	3
π_1 = 0.25	$\bar{R}_C^*$	.330	.177	.106	.335	.220	.088
	$\bar{R}_M^*$	.273	.152	.085	.270	.210	.081
	N_C	3	6	5	3	10	9
	N_M	21	14	12	19	14	9
π_1 = 0.35	$\bar{R}_C^*$	.337	.180	.107	.320	.240	.097
	$\bar{R}_M^*$	.320	.167	.092	.305	.226	.073
	N_C	6	7	4	7	7	7
	N_M	13	10	11	16	8	8
π_1 = 0.5	$\bar{R}_C^*$	.341	.169	.082	.330	.174	.086
	$\bar{R}_M^*$	.361	.210	.087	.340	.234	.099
	N_C	10	15	4	10	19	6
	N_M	6	2	2	7	3	4

Table 1. Error rates R_C^* and R_M^* for n = 40, p = 2.

		Sampling Scheme					
		Mixture			Separate		
			Δ			Δ	
		1	2	3	1	2	3
π_1 = 0.25	$\bar{R}_C$	.358	.223	.106	.360	.243	.114
	$\bar{R}_M$	.306	.194	.094	.300	.233	.092
π_1 = 0.35	$\bar{R}_C$	.349	.213	.102	.341	.253	.108
	$\bar{R}_M$	.348	.211	.091	.333	.244	.081
π_1 = 0.5	$\bar{R}_C$	.349	.189	.082	.351	.192	.085
	$\bar{R}_M$	.386	.230	.085	.378	.245	.098

tic search was conducted on each trial in an attempt to locate the global solution. Only results corresponding to $n=40$ and $p=2$ are reported here as the others exhibited similar behaviour.

3.1 Simulation results

The simulation results obtained are displayed in Tables 1 and 2. In Table 1 the average overall error rates $R_j^*(j=M, C)$ under the respective sampling schemes are listed for combinations $n=40$, $p=2$, $\Delta=1, 2, 3$, $\pi_1=.25$, $.35$, $.5$, along with the corresponding values of N_C and N_M, where N_C (N_M) is the number of times R_C^* (R_M^*) was less than R_M^* (R_C^*) out of the 30 trials simulated per combination of the parameters. The corresponding average over all error rates with respect to a subsequent observation, R_j $(j=M, C)$, is listed in Table 2. The results in Tables 1 and 2 suggest that whenever the mixing proportions or the sample sizes are equal or close to equality, CA is generally preferable to MA. The values of N_C and N_M in Tables 1 and 2 reveal that the superiority of the former approach over MA is more marked under separate than mixture sampling.

For mixing proportions which are far from 0.5 under the mixture sampling scheme, or for disparate sample sizes under the separate sampling scheme, it appears that MA is preferable in general to CA. The superiority of the former over the latter appears to be more marked under mixture than separate sampling. The tendency for CA to partition the sample into groups of approximately equal size explains the inferior performance of CA relative to MA in circumstances where the sample does not contain approximately the same number of observations from each subpopulation.

It was noted that for several combinations of the parameters, the discriminant function coefficients are estimated in roughly the same proportions under both procedures, and this explains the high proportions of ties between R_C^* and R_M^* as can be seen in Table 1.

In an extensive simulation study Ganesalingam and McLachlan (1979a, 1980) compared the mixture approach (MA) with the standard approach (SA), where it was observed that, though $\mathbf{a}_M$ and $\mathbf{b}_M$ are poor estimates of $\mathbf{a}$, and $\mathbf{b}$, in particular with small n, the ratio of $\mathbf{b}_M$ and $\mathbf{a}_M$ is always fairly close to the ratio of $\mathbf{b}_S$ and $\mathbf{a}_S$. Thus the coefficients $\mathbf{a}_M$, $\mathbf{b}_M$ are estimated in such proportions that W_M performs almost as well as W_S in terms of the overall error rate, even for n as small as 20 if the Mahalanobis distance between the populations is at least 2. The standard approach (SA) refers to the straightforward procedure of using maximum likelihood for completely classified data, where μ_i is estimated by the sample mean of the observations from Π_i; π_i is estimated by the proportion of observations from Π_i $(i=1, 2)$, and Σ is estimated by the pooled sample covariance matrix.

It is therefore recommended that under the normality model with equal covariance, MA should be used in preference to CA as a clustering procedure, unless one can be sure that the unclassified observations are present in approximately the same proportion from each subpopulation. In the latter instance CA outperforms MA as a clustering procedure. These recommendations hold regardless of whether the data were obtained by a mixture or a separate sampling scheme. However, where MA is applicable its superiority is more marked under mixture than separate sampling. Conversely, where CA is applicable its superiority is more marked under separate than mixture sampling. If the object is to use the unclassified sample to estimate the discriminant function coefficients, say, for use in forming posterior probabilities, then the MA is recommended in all instances.

4. Case study

We wish to test the validity of the above recommendations by applying the two methods to real data. They are compared not only on their ability to allocate the unclassified sample correctly but also on the bais of the estimates they give for the discriminant function coefficients, and hence for the posterior probabilities. The data to be analysed is taken from Habbema *et al.* (1974). In the context of genetic counselling, it was necessary to discriminate normal women and haemophilia A carriers on the basis of the two variables,

$X_1=\log_{10}$ (AHF activity)

$X_2=\log_{10}$ (AHF-like antigen).

Reference data were available from $n_1=30$ observations on known non-carriers and $n_2=45$ observations on known obligatory carriers. These data points are plotted in Fig. 1 along with the allocation boundaries

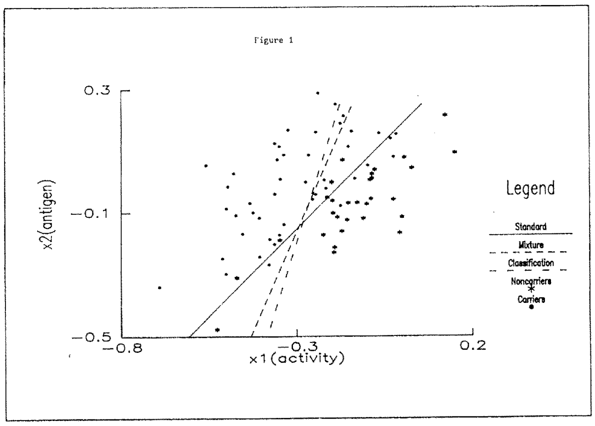

Figure 1. Haemophilia data with allocation boundaries. — SA, --- MA, ----- CA; *non-carriers and • carriers.

Table 3. Sample means and covariance matrices for $n_1 = 30$ noncarriers (Π_1) and $n_2 = 45$ carriers (Π_2).

Π_1		Π_2	
$\bar{x}_1$	S_1	$\bar{x}_2$	S_2
-.135	.021	-.308	.024
-.078	.018	-.006	.024
	.016		.015

to be discussed in the next section. Habbema *et al.* (1974) concluded from the reference data that the assumption of bivariate normal distribution with equal covariance matrices was reasonable. We let Π_1, be the subpopulation of noncarriers or normals and Π_2 be the subpopulation of carriers. The sample means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ and covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ for Π_1 and Π_2 are given in Table 3, where the diagonal elements of $\mathbf{S}_i$ are listed first followed by the off-diagonal elements (i = 1, 2).

In addition to the data of known classification as summarized in Table 3, there is also available from Habbema *et al.* (1974) a sample of 23 possible carriers $\mathbf{y}_k$ with known prior probabilities of noncarriership π_{1k} (k = 1, 2, ..., 23), which were the genetic chances of being normal deduced from their heredity. To highlight the differences that can occur with using MA and CA to estimate posterior probabilities for unclassified observations, we shall estimate the posterior probability that $\mathbf{y}_k$ is associated with a noncarrier, given under the normality assumption by

$$\theta_{1k} = 1/[1 + \exp\ \{\mathbf{a}'\ \mathbf{y}_k + \mathbf{b}^\circ + \log\ (\pi_{2k}/\pi_{1k})\}\ ].$$

4.1 Application of MA and CA

MA and CA are now applied to the classified bivariate data plotted in Fig. 1 as if it were an unclassified sample of n = 75 observations. Given that this sample has been formed by combining samples obtained separately on carriers and noncarriers, the mixture model does not hold. However, MA still appears to give reasonable results for separate sampling.

The maximum likelihood estimates of the discriminant function coefficients obtained under MA and CA are displayed in Table 4 along with the corresponding estimates for the standard approach, SA, based on the known classification of the sample. In Table 4 an estimate of b is not listed for the CA since with this approach allocation is performed always on the basis of

$$\mathbf{W}_C^\circ\ (\mathrm{x}) = \mathbf{a}_C'\,\mathbf{x} + \mathbf{b}_C^\circ;$$

that is, on the estimated log-likelihood ratio of the densities with a zero cut-off point rather than with a cut-off point depending on the estimated prior probabilities.

The allocation boundaries for MA and CA given by $W_M = 0$ and $W_C^\circ = 0$ respectively are plotted in Fig. 1, along with the boundary of the linear discriminant function $W_S = 0$, obtained using standard approach (SA). It can be observed that W_C° misallocates two more observations (both carriers) than W_M. Altogether, W_C° misallocates 3 and 15 observations from Π_1 and Π_2 respectively with W_M misallocating two fewer observations from Π_2. The overall error rates of 18/75 = 0.240 and 16/75 = 0.213 for W_C° and W_M respectively compare reasonably well with the corresponding error rate of 12/75 = 0.160 for W_S, considering that W_S is based on the correct classification of the sample, (W_S misallocates 5 and 7 observations from Π_1 and Π_2 respectively).

The discriminant function coefficients, $\mathbf{a}_C$ and $\mathbf{b}_C^\circ$ are not in close agreement with $\mathbf{a}_M$ and $\mathbf{b}_M^\circ$. Although the true values of a and b are not available, it would appear from a comparison with the standard estimates, $\mathbf{a}_S$ and $\mathbf{b}_S^\circ$, that $\mathbf{a}_C$ and $\mathbf{b}_C^\circ$ are appreciably under - estimating the components of **a and b**°, illustrating the bias problems that can occur with CA.

For this particular example where mixture sampling does not apply, one may want to place less emphasis on the estimated proportions by clustering the unclassified sample on the basis of

$$\mathbf{W}_M^\circ = \mathbf{a}_M'\mathbf{x} + \mathbf{b}_M^\circ, \tag{4.1.1}$$

Table 4. Estimates of discriminant function coefficients.

	Estimates of the Coefficients			
Approach	a_1	a_2	b^0	b
Mixture	-25.58	9.91	-5.94	-6.05
Classification	-38.00	10.54	-9.31	
Standard	-19.34	17.13	-3.57	-3.17

Table 5. Data, prior probabilities and estimated posterior probabilities of non-carriership for 23 possible haemophilia A carriers.

Observation number	Data		Prior probability of non-carriership	Approach		
				Classification	mixture	Standard
k	y_{1k}	y_{2k}	π_{1k}	θ^C_{1k}	θ^M_{1k}	θ^S_{1k}
1	-0.112	-0.279	0.50	1.000	0.997	0.998
2	-0.059	-0.068	0.75	1.000	0.999	0.991
3	0.064	0.012	0.50	1.000	1.000	0.990
4	-0.043	-0.052	0.67	1.000	0.991	0.987
5	-0.050	-0.098	0.50	1.000	0.999	0.986
6	-0.094	-0.113	0.50	0.999	0.991	0.976
7	-0.123	-0.143	0.34	0.996	0.985	0.951
8	-0.011	0.037	0.50	1.000	0.995	0.938
9	-0.210	-0.090	0.67	0.951	0.896	0.851
10	-0.126	-0.019	0.50	0.991	0.948	0.810
11	-0.210	-0.044	0.75	0.947	0.891	0.794
12	0.069	0.192	0.34	1.000	0.999	0.720
13	0.030	0.224	0.50	1.000	0.979	0.576
14	-0.002	0.206	0.50	0.999	0.960	0.499
15	-0.371	-0.133	0.50	0.032	0.096	0.208
16	-0.162	0.162	0.34	0.682	0.380	0.046
17	-0.097	0.263	0.34	0.899	0.701	0.031
18	-0.250	0.095	0.34	0.134	0.197	0.028
19	-0.287	0.137	0.50	0.046	0.059	0.013
20	-0.357	0.031	0.34	0.005	0.015	0.011
21	-0.461	-0.003	0.50	0.000	0.002	0.005
22	-0.590	-0.139	0.50	0.000	0.000	0.004
23	-0.561	-0.060	0.50	0.000	0.000	0.002

which corresponds to using the estimated log likelihood ratio with a zero cut-off point. In this instance as $\pi_{1M}=0.529$, b°_M is very similar to b_M (see Table 4), and using (4.1.1) gives the same allocation as W_M applied to the $n=75$ observations in Fig. 1. However, since the estimates $\mathbf{a}_M$ and $\mathbf{b}^\circ_M$ have been obtained under the assumption that the unclassified sample has been drawn from a mixture of two subpopulations in some proportions π_1 and π_2, then there is good reason to cluster the sample on the basis of this assumption, using W_M as the discriminant function where the cut-off point depends on the estimated proportions.

The effect of using $\mathbf{b}^\circ_M$ instead of $\mathbf{b}_M$ as the constant term in the discriminant function is to produce clusters of roughly equal size which we have seen with CA is not a desirable feature if the sample happens to contain disparate numbers of observations from the subpopulations. If a zero cut-off point is felt desirable, then perhaps $\mathbf{a}$ and $\mathbf{b}^\circ$ should be estimated with the mixing proportions taken to be equal during the iterative computations under MA.

4.2 Estimates of posterior probabilities

We now consider the estimated posterior probability of belonging to Π_1 (that is, of being a non-carrier) for each of the 23 possible carriers whose observation vectors $\mathbf{y}_k$ and known prior probabilities π_{1i} were given in Table 5 along with the various estimates obtained for each θ_{1k}, the posterior probability that $\mathbf{y}_k$ belongs to Π_1 where θ^M_{1k}, θ^C_{1k} and θ^S_{1k} denote the estimates of θ_{1k} using the mixture, classification, and standard approaches respectively to estimate $\mathbf{a}$ and $\mathbf{b}^\circ$. The observations $\mathbf{y}_k$ have been labelled so as to be in order of decreasing size of θ^S_{1k}. We see from Table 5 that, except for the $\mathbf{y}_k$ for which $\theta^C_{1k}=\theta^M_{1k}$, the CA gives a more extreme estimate of θ_{1k} with θ^M_{1k} being nearer to 0.5. Hence the bias associated with the CA manifests itself in the estimates of the posterior probabilities.

Except for $\mathbf{y}_{16}$, CA allocates the $\mathbf{y}_k$ to either Π_1 or Π_2 in the same pattern as MA, which gives a similar allocation to that of the standard approach (SA). The exceptions are $\mathbf{y}_{17}$ and $\mathbf{y}_{14}$. $\mathbf{y}_{14}$ is just allocated to Π_2 by SA (it is very near to the boundary) but is decisively allocated to Π_1 by MA. Some idea of the reliability of the estimates of the posterior probabilities provided by CA and MA can be made by comparing θ^C_{1k} and θ^M_{1k} with θ^S_{1k}. For $k=1, 2, \ldots, 8$ and $k=16, \ldots, 23$, $\mathbf{y}_k$ has either a very high or low values of θ^S_{1k}, and so there is little doubt as to which subpopulation $\mathbf{y}_k$ belongs. For a majority of these cases the CA and MA give a more extreme estimate of θ_{1k}. For k = 10, ..., 15, $\mathbf{y}_k$ has a more doubtful estimate of population membership according to θ^{S}_{1k}; indeed, θ^{S}_{1k} is close to 0.5 for $k = 13$ and 14. Yet CA and MA give estimates of θ_{1k} close to one. Further disagreement between the θ^{S}_{1k} and θ^{C}_{1k} or θ^{M}_{1k} can be seen for k = 16, 17 and 18. Overall, it can be seen that MA and CA give very misleading estimates of θ_{1k} relative to θ^{S}_{1k}, at least for k = 12, 13, 14, 16, and 17.

Hence for a sample size of n = 75 it appears that the reliability of the estimated posterior probabilities provided by CA and MA are very questionable if there is any doubt associated with subpopulation membership; that is, if the value of the posterior probability is in the interval (0.1, 0.9). This is unfortunate as it is precisely such situations where reliable estimates of subpopulation memberships are needed. If more reliable estimates are required, then the sample size n must be very large. It has been noted previously (Day 1969, Hosmer 1973 and Ganesalingam 1980), that *n* has to be very large for MA to give good estimates of the subpopulation parameters.

4.3 Estimation of the overall error rate using plug-in and posterior probability based estimators

For the case study considered in Section 4 it can be established from the known classification of the sample that the apparent error rate of W_M applied to the sample of n = 75 observations is equal to

$$R^*_M = 16/75 = 0.213$$

We shall now use the two well known estimators namely plug-in and posterior probability based estimators to estimate the overall error rate associated with W_M formed from the sample of n = 75 bivariate observations. That is, by treating this sample as if it were unclassified and not making use of its known classification we are to estimate the overall error rate of W_M which also has been formed without this knowledge.

These estimators are defined as follows. The posterior probability based estimator e is given by

$$e = \sum_{k=1}^{n} \min\{\hat{\theta}_1(\mathbf{z}_k), \hat{\theta}_2(\mathbf{z}_k)\}/n$$

where $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ denote n observations drawn from the mixture of Π_1 and Π_2 in the proportions π_1 and π_2 and $\theta_i(\mathrm{z}_k)$ denotes the posterior probability of $\mathbf{z}_k$ belonging to Π_i (i = 1, 2). The bias of **e** was examined by deriving an asymptotic approximation, Ganesalingam (1980).

The plug-in estimator is given by

$$\mathbf{Q} = \pi_{1M}\,\mathbf{Q}_1 + \pi_{2M}\,\mathbf{Q}_2$$

where

$$\mathbf{Q}_i = \Phi[\{-\frac{1}{2}\,\hat{\Delta}^2_M + (-1)^i \log(\pi_{1M}/\pi_{2M})\}/\hat{\Delta}_M]$$

and where

$$\hat{\Delta}^2_M = \mathbf{a}'_M(\hat{\mu}_{2M} - \hat{\mu}_{1M}).$$

The posterior probability based estimator e has an advantage over the plug-in estimator Q in that it can easily make use of further unclassified observations without having to recompute the estimates of the parameters.

It was noticed by Ganesalingam (1980) that the bias associated with **e** and **Q** are not small in magnitude with **e** and **Q** appreciably underestimating the overall error rate associated with W_M, R_M if n is not large relative to p. Hence some method of bias correction such as jackknife procedure (Quenouille 1956) or bootstrap procedure of Efron (1979) should be considered with application of **e** and **Q** in such instances although it may involve considerable computation.

The estimates of R^*_M given by **e** and **Q** are displayed in Table 6, along with their bias corrected forms obtained by using the bootstrap and jackknife procedures. For this example, the bootstrap version of the plug-in estimator gives the closer estimate, 0.232 to the apparent error rate, $R^*_M = 0.213$. Regarding the estimation of the overall error rate of W_M applied to a subsequent observation, R_M, the bootstrap versions of **e** and **Q** yield 0.253 and 0.244 respectively as the estimates of R_M. A question worth considering in any future research into this error rate estimation problem on the basis of an unclassified sample is how the boostrap and jackknife methods compare in general in terms of the criteria such as bias and mean square error.

Acknowledgements. I thank Professor R. Mead of the University of Reading and my colleague Mr G.C. Arnold for a number of comments and suggestions which helped me to improve the presentation.

Table 6. Estimates of the overall apparent error rate, R^*_M of W_M ($R^*_M = 0.213$).

		Methods of bias correction	
Type of Estimator	**Uncorrected**	**Bootstrap**	**Jackknife**
Posterior Probability	0.105	0.241	0.265
Plug-in	0.109	0.232	0.247

REFERENCES

Andrews, D.F. 1973. Graphical techniques for high dimensional data. *In:* Cacoullous, T. (ed.). *Discriminant Analysis and Application*, pp. 37-59.

Bryant, P. and J.A. Williamson. 1978. Asymptotic behaviour of classification maximum likelihood estimates. Biometrika 65: 273-281.

Day, N.E. 1969. Estimating the components of a mixture of two normal distributions. Biometrika 56: 463-474.

Dempster, A.P., N.M. Laird and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39: 1-38.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7: 1-26.

Ganesalingam, S. 1980. *On the Mixture Maximum Likelihood Approach to Estimation and Clustering.* Unpublished Ph. D. thesis, University of Queensland, Australia.

Ganesalingam, S. and G.J. McLachlan. 1978. The efficiency of a linear discriminant function based on unclassified initial samples. Biometrika 65: 658-662.

Ganesalingam, S. and G.J. McLachlan. 1979. Small sample results for a linear discriminant function estimated from a mixture of normal populations. Journal of Statistical Computation and Simulation 9: 151-158.

Ganesalingam, S. and G.J. McLachlan. 1980. A comparison of the mixture and classification approaches to cluster analysis. Communications in Statistics - Theory and Methods A9: 923-933.

Gnanadesikan, R. 1977. *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley, New York.

Habbema, J.D.F., J. Hermans and K. Van Den, Broek. 1974. A stepwise discriminant analysis program using density estimation. Compstat 1974 Proceedings in Computational Statistics, pp. 101-110. Physica Verlag, Wien.

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

Hartigan, J.A. 1978. Asymptotic distributions for clustering criteria. The Annals of Statistics 6: 117-131.

Hosmer, D.W. 1973. On MLE of the parameters of a mixture of two normal distributions when the sample size is small. Communications in Statistics 1: 217-227.

John, S. 1970a. On identifying the population of origin of each observation in a mixture of observations from two normal populations. Technometrics 12: 553-563.

John, S. 1970b. On identifying the population of origin of each observation in a mixture of observations from two gamma populations. Technometrics 12: 565-568.

MacQueen, J. 1966. Some methods for classification and analysis of multivariate observations. Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability 1: 281-297.

Marriott, F.H.C. 1971. Practical problems in a method of cluster analysis. Biometrics 27: 501-514.

Marriott, F.H.C. 1975. Separating mixtures of normal distributions. Biometrics 31: 767-769.

McLachlan, G.J. 1980. The classification and maximum likelihood approaches to cluster analysis. *In:* Krishnaiah, P.R. and L.N. Kanal (eds.). *Handbook of Statistics*, Vol. 2, pp. 109-208. North-Holland, Amsterdam.

O'Neill, T.J. 1978. Normal discrimination with unclassified observations. Journal of the American Statistical Association 73: 821-826.

Quenouille, N. 1956. Notes on bias in estimation. Biometrika 43: 353-360.

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

Manuscript received: December 1988
Final revision October 1989