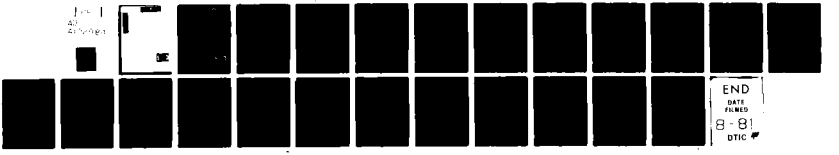
AD-A102 084 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
THE CLASSIFICATION AND MIXTURE MAXIMUM LIKELIHOOD APPROACHES TO--ETC(U)
MAR 81 G J MCLACHLAN N00014-76-C-0475
UNCLASSIFIED TR-299 NL



END
DATE FILMED
8-81
DTIC

(11) LEVEL II

THE CLASSIFICATION AND MIXTURE MAXIMUM LIKELIHOOD APPROACHES TO CLUSTER ANALYSIS.

By

(10) G. J. McLACHLAN

11 12 Mar 81

(9) TECHNICAL REPORT NO. 299

MARCH 12, 1981

(11) TR-299

(12)

Prepared Under Contract
(15) N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATSITICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis*

G.J. McLachlan

1. INTRODUCTION

A common and very old problem in statistics is the separation of a heterogeneous population into more homogeneous subpopulations. We concentrate here on the situation where the population of interest, Π, is known or assumed to consist of, say, k different subpopulations $\Pi_1,\ldots,\Pi_k$, and where the density of a p-dimensional observation $\underset{\sim}{x}$ from Π_i is known or assumed to be $f_i(\underset{\sim}{x};\underset{\sim}{\theta})$ for some unknown vector of parameters, $\underset{\sim}{\theta}$ $(i=1,\ldots,k)$. In this context the problem may be formulated as follows: Given a random sample of observations $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$ from Π, attempt to allocate each $\underset{\sim}{x}_j$ to the subpopulation to which it belongs. We let $\underset{\sim}{\gamma}' = (\gamma_1,\ldots,\gamma_n)$ denote the set of identifying labels, where $\gamma_j = i$ if $\underset{\sim}{x}_j$ comes from Π_i. This would be the classical discrimination problem if $\underset{\sim}{\gamma}$ were known a priori; a discrimination procedure would be formed from the classified sample for the allocation of subsequent observations of unknown origin.

In what is sometimes called the classification maximum likelihood procedure, $\underset{\sim}{\theta}$ and $\underset{\sim}{\gamma}$ are chosen to maximize

$$L_C(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n;\underset{\sim}{\theta},\underset{\sim}{\gamma}) = \prod_{j=1}^{n} f_{\gamma_j}(\underset{\sim}{x}_j;\underset{\sim}{\theta}) . \tag{1.1}$$

The maximization is over the set of values of $\underset{\sim}{\gamma}$ corresponding to all possible assignments of the $\underset{\sim}{x}_j$ to the various subpopulations as well as over all admissible values of $\underset{\sim}{\theta}$. The estimates of $\underset{\sim}{\theta}$ and $\underset{\sim}{\gamma}$ so obtained are denoted by $\hat{\underset{\sim}{\theta}}$ and $\hat{\underset{\sim}{\gamma}}$ respectively. The

*To appear in Vol. II of the Handbook of Statistics (edited by P.R. Krishnaiah and L. Kanal).

$\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$ are then classified according to the estimates $\tilde{\gamma}_1,\ldots,\tilde{\gamma}_n$; for example, $\underset{\sim}{x}_j$ is assigned to Π_g if $\tilde{\gamma}_j = g$. This procedure has been considered by several authors including Hartley and Rao [14], John [17], Scott and Symons [31], and Sclove [30]. Unfortunately, with this procedure, the γ_j increase in number with the number of observations, and under such conditions the maximum likelihood estimates need not be consistent. Marriott [23] pointed out that under the standard assumption of normal distributions with common variance matrices, this procedure gives definitely inconsistent estimates for the parameters involved. More recently, Bryant and Williamson [4] extended Marriott's results and showed that the method may be expected to give asymptotically biased results quite generally.

A related approach is the mixture maximum likelihood method considered by Day [5], and Wolfe [34], among many others. With this approach $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$ are assumed to be a random sample of size n from a mixture of $\Pi_1,\ldots,\Pi_k$ in the proportions $(\varepsilon_1,\ldots,\varepsilon_k) = \underset{\sim}{\varepsilon}'$. Hence the likelihood

$$L_M(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n;\underset{\sim}{\theta},\underset{\sim}{\varepsilon}) = \prod_{j=1}^{n} \{\sum_{i=1}^{k} \varepsilon_i f_i(\underset{\sim}{x}_j;\underset{\sim}{\theta})\} \tag{1.2}$$

can be formed; the estimates of $\underset{\sim}{\theta}$ and $\underset{\sim}{\varepsilon}$ obtained by maximizing (1.2) are denoted by $\hat{\underset{\sim}{\theta}}$ and $\hat{\underset{\sim}{\varepsilon}}$ respectively. Each $\underset{\sim}{x}_j$ can be classified then on the basis of the estimated posterior probabilities $\hat{P}_{ij}$ $(i=1,\ldots,k)$ formed by replacing $\underset{\sim}{\theta}$ and $\underset{\sim}{\varepsilon}$ with $\hat{\underset{\sim}{\theta}}$ and $\hat{\underset{\sim}{\varepsilon}}$ in

$$P_{ij} = \Pr\{\underset{\sim}{x}_j \in \Pi_i | \underset{\sim}{x}_j\} \quad ,$$

and $\underset{\sim}{x}_j$ is assigned to Π_g if

$$\hat{P}_{gj} \geq \hat{P}_{ij} \quad (i=1,\ldots,k) \; .$$

It can be seen that the mixture approach is equivalent to the classification procedure with the additional assumption that $\gamma_1,\ldots,\gamma_n$ is an (unobservable) random sample from a probability distribution with mass ε_i at i $(i=1,\ldots,k)$. It appears to avoid the asymptotic biases associated with the classification procedure where at each step in the iterative process of computing the maximum likelihood estimates each $\underset{\sim}{x}_j$ is assigned outright to a particular subpopulation according to the estimate for γ_j. By contrast, the mixture approach does not insist on definite membership to any subpopulation; rather it gives an estimated probability of membership of each subpopulation.

Note that another approach to this problem is to proceed further and adopt a Bayesian procedure in which all parameters are random variables (Binder [2], Symons [32]).

A common assumption in practice is to adopt the normality model

$$\underset{\sim}{x}_j \sim N(\underset{\sim}{\mu}_i, \underset{\sim}{\Sigma}) \quad \text{in} \quad \Pi_i \quad (i=1,\ldots,k) \; . \tag{1.3}$$

In this case $\underset{\sim}{\theta}$ has $\frac{1}{2}p(p+2k+1)$ elements, comprising the components

of the k mean vectors $\underset{\sim}{\mu}_i$ and the distinct elements of the common covariance matrix $\underset{\sim}{\Sigma}$, and the density $f_i(\underset{\sim}{x};\underset{\sim}{\theta})$ is given by

$$f(\underset{\sim}{x};\underset{\sim}{\mu}_i,\underset{\sim}{\Sigma}) = (2\pi)^{-1/2\,p}\,|\underset{\sim}{\Sigma}|^{-1/2}\{\exp -\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_i)'\,\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_i)\}\ .$$

We now proceed to consider the application of the classification and mixture approaches under the normality model (1.3) which is assumed to hold through to Section 5, where the condition of a common covariance matrix is relaxed to cover the general case of unequal covariance matrices.

2. CLASSIFICATION APPROACH

In principle the maximization process for the classification maximum likelihood procedure can be carried out since it is just a matter of computing the maximum value of the likelihood (1.1) over all possible partitions of the n observations to the k subpopulations. However, unless n is quite small, searching over all possible partitions is prohibitive. It follows that $\tilde{\gamma}_j = g$ if

$$f(\underset{\sim}{x}_j;\tilde{\underset{\sim}{\mu}}_g,\tilde{\underset{\sim}{\Sigma}}) \geq f(\underset{\sim}{x}_j;\tilde{\underset{\sim}{\mu}}_i,\tilde{\underset{\sim}{\Sigma}}),\quad (i=1,\ldots,k)\ , \tag{2.1}$$

where $\tilde{\underset{\sim}{\mu}}_i$ and $\tilde{\underset{\sim}{\Sigma}}$ are the ordinary maximum likelihood estimates of $\underset{\sim}{\mu}_i$ and $\underset{\sim}{\Sigma}$ for a sample of normal observations classified according to $\tilde{\underset{\sim}{\gamma}}$. Hence the solution can be computed iteratively (John [17], Sclove [30]). Starting with some initial clustering $\underset{\sim}{\gamma}$, the $\underset{\sim}{\mu}_i$

and $\underset{\sim}{\Sigma}$ are estimated accordingly and then used to give a new estimate of $\underset{\sim}{\gamma}$ on the basis (2.1), equivalent to allocating each observation to the nearest cluster centre in terms of the estimated Mahalanobis distance. Each step in the iterative process yields a value of the likelihood not less than that at the previous step, and the iterations may be continued until no observation changes clusters. Various starting values should be taken in an attempt to locate the global solution. It will be seen in the next section that the likelihood equations under the mixture approach can be easily modified to be applicable also under the classification approach. There are other procedures for finding the solution under the classification approach; for example, the Mahalanobis distance version of MacQueen's [20] k-means procedure, where the $\underset{\sim}{\mu}_i$ and $\underset{\sim}{\Sigma}$ are re-estimated after each observation is allocated rather than waiting until after all the observations have been allocated.

For the classification approach applied under the normality model (1.3), Scott and Symons [31] showed that $\tilde{\underset{\sim}{\gamma}}$ corresponds to the partition which minimizes the determinant of the pooled within-subpopulations sum of squares matrix

$$\underset{\sim}{W} = \sum_{i=1}^{k} \underset{\sim}{W}_i ,$$

where

$$\underset{\sim}{W}_i = \sum_{q=1}^{n_i} (\underset{\sim}{x}_{iq} - \bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_{iq} - \bar{\underset{\sim}{x}}_i)'$$

and $\underset{\sim}{x}_{iq}$ $(q=1,\ldots,n_i)$ denote the n_i observations assinged to Π_i

according to $\tilde{\gamma}$ and $\bar{x}_i$ refers to their sample mean; see also Friedman and Rubin [9] who originally suggested this criterion. The minimization of $|W|$ would appear to be a reasonable clustering criterion regardless of the underlying distributions. Marriott [22] has given a comprehensive account of the properties of this criterion. It does have the tendency to produce clusters of roughly equal size, although the modified version,

$$n \log|W| - 2 \sum_{i=1}^{k} n_i \log n_i$$

suggested recently by Symons [32], would appear to go some way to overcoming this.

3. MIXTURE APPROACH

An excellent account of the computation of the maximum likelihood estimates of μ_i, Σ, and ε for the mixture approach has been given by Day [5]. Under the normality model (1.3), the posterior probabilities $P_{ij}(i=1,\ldots,k;j=1,\ldots,n)$ have the form

$$P_{ij} = \exp(a_i' x_j + b_i) / \{ \sum_{r=1}^{k} \exp(a_r' x_j + b_r) \}$$

where

$$a_r = \Sigma^{-1}(\mu_r - \mu_1)$$

and

$$b_r = \frac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_r)' \underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_r) + \log(\varepsilon_r/\varepsilon_1)$$

for $r = 1,\ldots,k$; that is, $\underset{\sim}{a}_1 = \underset{\sim}{0}$ and $b_1 = 0$. The maximum likelihood estimates are evaluated from the equations

$$\hat{\varepsilon}_i = \sum_{j=1}^{n} \hat{P}_{ij}/n \tag{3.1}$$

$$\hat{\underset{\sim}{\mu}}_i = \sum_{j=1}^{n} (\hat{P}_{ij}\underset{\sim}{x}_j)/(n\,\hat{\varepsilon}_i) \tag{3.2}$$

and

$$\hat{\underset{\sim}{\Sigma}} = \sum_{i=1}^{k}\sum_{j=1}^{n} (\hat{P}_{ij}/n)(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)' , \tag{3.3}$$

which can be solved iteratively by substituting some initial values for the estimates into the right-hand side of (3.1) to (3.3) to produce new estimates on the left-hand side, which are then substituted into the right-hand side, and so on. These iterative estimates can be identified with those obtained by directly applying the so-called EM algorithm of Dempster et al. [6], which shows that the estimates will converge to a local maximum irrespective of the starting point. The iterative process should be started from several points in an attempt to ensure that the global maximum is obtained.

Day [5] has shown that considerable computing time can be saved for $k = 2$ by reparametrizing the likelihood in terms of $\underset{\sim}{a}$, b, $\underset{\sim}{m}$, and $\underset{\sim}{V}$, where

$$\underset{\sim}{m} = \varepsilon_1\underset{\sim}{\mu}_1 + \varepsilon_2\underset{\sim}{\mu}_2$$

and

$$\underset{\sim}{V} = \underset{\sim}{\Sigma} + \varepsilon_1\varepsilon_2(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'$$

and the mean and covariance matrix of the mixture distribution; $\underset{\sim}{a}$ and b denote $\underset{\sim}{a}_2$ and b_2 with their subscripts suppressed since $k = 2$ only. The maximum likelihood equations now can be written as

$$\hat{\underset{\sim}{m}} = \sum_{j=1}^{n} \underset{\sim}{x}_j/n \ , \tag{3.4}$$

$$\hat{\underset{\sim}{V}} = \sum_{j=1}^{n} (\underset{\sim}{x}_j - \hat{\underset{\sim}{m}})(\underset{\sim}{x}_j - \hat{\underset{\sim}{m}})'/n \ , \tag{3.5}$$

$$\hat{\underset{\sim}{a}} = \hat{\underset{\sim}{V}}^{-1}(\hat{\underset{\sim}{\mu}}_2 - \hat{\underset{\sim}{\mu}}_1)/\{1-\hat{\varepsilon}_1\hat{\varepsilon}_2(\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)'\hat{\underset{\sim}{V}}^{-1}(\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)\} \tag{3.6}$$

and

$$\hat{b} = -\frac{1}{2}\hat{\underset{\sim}{a}}'(\hat{\underset{\sim}{\mu}}_1 + \hat{\underset{\sim}{\mu}}_2) + \log(\hat{\varepsilon}_2/\hat{\varepsilon}_1) \ . \tag{3.7}$$

Only values of $\hat{\underset{\sim}{a}}$ and $\hat{b}$ are needed in solving the above equations as $\hat{\underset{\sim}{m}}$ and $\hat{\underset{\sim}{V}}$ are given explicitly.

To obtain suitable initial values of $\underset{\sim}{a}$ and b, it is suggested for various bivariate subsets of the variables plotting the data points and drawing a line which divides the data into two groups which have a scatter that appears normal (see, for example, O'Neill [28] and Ganesalingam and McLachlan [12]). Estimates of $\underset{\sim}{a}$ and b can be formed on the basis of this subdivision, proceeding as if the observations were correctly classified. There appears to be no difficulty in locating the global maximum for $p = 1$ and 2, but for $p \geq 3$ there are problems with multiple maxima, particularly for small values (less than two, say) of the Mahalanobis distance between Π_1 and Π_2,

$$\Delta = \{(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\}^{1/2} ,$$

when n is not large (Day [5]). Also, it is well-known (Day [5] and Hosmer [16]) that maximum likelihood estimates based on a mixture of normal distributions are very poor unless n is very large (for example, $n \geq 500$). However, Ganesalingam and McLachlan [11] found that although the maximum likelihood estimates $\hat{\underset{\sim}{a}}$ and $\hat{b}$ may not be very reliable for small n, it appears that the proportions in which the components of $\hat{\underset{\sim}{a}}$ and $\hat{b}$ occur are such that the resulting discriminant function, $\hat{\underset{\sim}{a}}'\underset{\sim}{x} + \hat{b}$, may still provide reasonable separation between the subpopulations.

Note that the same set of equations here can be used as follows to compute the estimates $\tilde{\underset{\sim}{\mu}}_i$, $\tilde{\underset{\sim}{\Sigma}}$, and $\tilde{\underset{\sim}{\gamma}}$ under the classification approach.

At a given step $\tilde{\gamma}_j$ is put equal to that g for which $\tilde{P}_{gj} \geq \tilde{P}_{ij}$ $(i=1,\ldots,k)$ where, in the P_{ij}, b_r is used without the $\log(\varepsilon_r/\varepsilon_1)$ term. Then on the next step the $\tilde{\mu}_i$ and $\tilde{\Sigma}$ are computed from (3.1) to (3.3) in which, for each j, $\hat{P}_{ij}$ is replaced by 1 $(i=g)$ and 0 $(i\neq g)$. The transformed equations (3.4) to (3.7) for $k=2$ are also applicable to the classification approach with the above modifications; that is, the term corresponding to $\hat{\varepsilon}_i$ in (3.6) is given by n_i/n $(i=1,2)$ while there is no term corresponding to $\log(\hat{\varepsilon}_2/\hat{\varepsilon}_1)$ in (3.7).

A simulation study undertaken by Ganesalingam and McLachlan [13] for $k=2$ suggests that overall the mixture approach performs quite favourably relative to the classification approach even where mixture sampling does not apply. The apparent slight superiority of the latter approach for samples with subpopulations represented in approximately equal numbers is more than offset by its inferior performance for disparate representations.

4. EFFICIENCY OF THE MIXTURE APPROACH

We consider now the efficiency of the mixture approach for $k=2$ normal subpopulations, contrasting the asympotic theory with small sample results available from simulation.

For a mixture of two univariate normal distributions Ganesalingam and McLachlan [10] studied the asymptotic efficiency of the mixture approach relative to the classical discrimination procedure (appropriate for known γ) by considering the ratio

$$e = \{E(R) - R_o\}/\{E(R_M) - R_o\} , \tag{4.1}$$

where $E(R_M)$ and $E(R)$ denote the unconditional error rate of the mixture and classical procedures respectively applied to an unclassified observation subsequent to the inital sample, and R_o denotes their common limiting value as $n \to \infty$. The asymptotic relative efficiency was obtained by evaluating the numerator and denominator of (4.1) up to and including terms of order $1/n$. The multivariate analogue of this problem was considered independently by O'Neill [28]. By definition the asymptotic relative efficiency does not depend on n, and O'Neill [28] showed that it also does not depend on p for equal prior probabilities, $\epsilon_1 = 0.5$. The asymptotic values of e are displayed in Table 1 as percentages for selected combinations of Δ^2, ϵ_1, p, and n; the corresponding values of e obtained from simulation are extracted from Ganesalingam and McLachlan [11] and listed below in parentheses. It can be seen that the asymptotic relative efficiency does not give a reliable guide as to the true relative efficiency when n is small, particularly for $\Delta = 1$. This is not surprising since the asymptotic theory of maximum likelihood for this problem requires n to be very large before it applies (Day [5], Hosmer [16]). Further simulation studies by Ganesalingam and McLachlan [11] in the univariate case indicate that the asymptotic relative efficiency gives reliable predictions at least for $n \geq 100$ and $\Delta \geq 2$.

The simulated values for the relative efficiency in Table 1 suggest that for the mixture approach to perform comparably with the classical discrimination procedure it needs to be based on about two to five times the number of initial observations, depending on the combination of the parameters.

5. UNEQUAL COVARIANCE MATRICES

For normal subpopulations Π_i with unequal covariance matrices $\underset{\sim}{\Sigma}_i$, the classification procedure has to be applied with the restriction that at least $p+1$ observations belong to each subpopulation to avoid the degenerate case of infinite likelihood.

The likelihood equations under the mixture approach are given by (3.1) to (3.3) appropriately modified to allow for k different covariance matrices (Wolfe [34]). Unfortunately, maximum likelihood estimation breaks down in practice for each data point gives rise to a singularity in the likelihood on the edge of the parameter space. This problem has received a good deal of attention recently. For a mixture of two univariate normal distributions, Kiefer [18] has shown that the likelihood equations have a root $\hat{\underset{\sim}{\phi}}$ which is a consistent, asymptotically normal and efficient estimator of $\underset{\sim}{\phi} = (\underset{\sim}{\theta}', \underset{\sim}{\xi}')'$. Quandt and Ramsey [29] proposed the moment generating function (MGF) estimator obtained by minimizing

$$\sum_{i=1}^{h} \{\psi(t_i) - \sum_{j=1}^{n} e^{t_i x_j}/n\}^2$$

for selected values $t_1,\ldots,t_h$ of t in some small interval (c,d), $c < 0 < d$, where

$$\psi(t) = \sum_{i=1}^{2} \varepsilon_i \exp(\mu_i t + \frac{1}{2}\sigma_i^2 t^2)$$

is the MGF of a mixture of two normal distributions with variances σ_1^2 and σ_2^2. The usefulness of the MGF method would appear to be that it provides a consistent estimate which can be used as a starting value when applying the EM algorithm in an attempt to locate the root of the likelihood equations corresponding to the consistent, asymptotically efficient estimator. Bryant [3] suggests taking the classification maximum likelihood estimate of $\underset{\sim}{\phi}$ as a starting value in the likelihood equations.

The robustness of the mixture approach based on normality as a clustering procedure requires investigation. A recent case study by Hernandez-Alvi [15] suggests that, at least in the case where the variables are in the form of proportions, the mixture approach may be reasonably robust from a clustering point of view of separating samples in the presence of multimodality.

6. UNKNOWN NUMBER OF SUBPOPULATIONS

Frequently with the application of clustering techniques there is the difficult problem of deciding how many subpopulations, k, there are. A review of this problem has been given by Everitt [8]; see also

Engelman and Hartigan [7] and Lee [19]. With respect to the classification approach Marriott [21] has suggested taking k to be the number which minimizes $k^2|\underset{\sim}{W}|$. For heterogeneous covariance matrices there may be some excessive subdivision, but this can be rectified by recombining any two clusters which by themselves do not suggest separation was necessary.

With the mixture approach the likelihood ratio test is an obvious criterion for choosing the number of subpopulations. However, for testing the hypothesis of, say, k_1 versus k_2 subpopulations ($k_1 < k_2$), it has been noted (Wolfe [35]) that some of the regularity conditions are not satisfied for minus twice the log-likelihood ratio to have under the null hypothesis an approximate chi-square distribution with degrees of freedom equal to the difference in the number of parameters in the two hypotheses. Wolfe [35] suggested using a chi-square distribution with twice the difference in the number of parameters (not including the proportions), which appears to be a reasonable approximation (Hernandez-Alvi [15]).

7. PARTIAL CLASSIFICATION OF SAMPLE

We now consider the situation where the classification of some of the observations in the sample is initially known. This information can be easily incorporated into the maximum likelihood procedures for the classification and mixture approaches. If an $\underset{\sim}{x}_j$ is known to come from, say Π_r, then under the former approach $\gamma_j = r$ always in the associated iterative process while, under the latter, P_{ij} is set equal to $1(i = r)$

and $O(i \neq r)$ in all the iterations. In those situations where there are sufficient data of known classification to form a reliable discrimination rule, the unclassified data can be clustered simply according to this rule and, for the classification approach, the results of McLachlan [24,25] suggest this may be preferable unless the unclassified data are in approximately the same proportion from each subpopulation. With the mixture approach a more efficient clustering of the unclassified observations should be obtained by simultaneously using them in the estimation of the subpopulation parameters, at least as $n \to \infty$, since the procedure is asymptotically efficient. The question of whether it is a worthwhile exercise to update a discrimination rule on the basis of a limited number of unclassified observations has been considered recently by McLachlan and Ganesalingam [26]. For other work on the updating problem the reader is referred to Titterington [33], Murray and Titterington [27], and Anderson [1].

ACKNOWLEDGEMENT

This work was completed while the author was on leave with the Department of Statistics at Stanford University.

TABLE 1

Asymptotic Versus Simulation Results for the Relative Efficiency of the Mixture Approach

	p=1 , n=20		p=2 , n=20		p=3 , n=40	
Δ	$\varepsilon_1 = 0.25$	$\varepsilon_1 = 0.50$	$\varepsilon_1 = 0.25$	$\varepsilon_1 = 0.50$	$\varepsilon_1 = 0.25$	$\varepsilon_1 = 0.50$
1	0.25	0.51	0.34	0.51	0.42	0.51
	(33.01)	(25.12)	(46.71)	(63.11)	(25.00)	(43.39)
2	7.29	10.08	9.36	10.08	10.51	10.08
	(22.05)	(17.74)	(25.73)	(16.26)	(16.28)	(14.51)
3	31.41	35.92	35.13	35.92	36.78	35.92
	(19.57)	(23.54)	(43.91)	(29.63)	(29.01)	(23.46)

REFERENCES

[1] Anderson, J.A. (1979). Multivariate logistic compounds. Biometrika, 66, 7-16.

[2] Binder, D.A. (1978). Bayesian cluster analysis. Biometrika, 65, 31-38.

[3] Bryant, P. (1978). Contribution to the discussion of the paper by R.E. Quandt and J.B. Ramsey. Journal of the American Statistical Association, 73, 748-749,

[4] Bryant, P. and Williamson, J.A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. Biometrika, 65, 273-281.

[5] Day, N.E. (1969). Estimating the components of a mixture of normal distributions. Biometrika, 56, 463-474.

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

[7] Engelman, L. and Hartigan, J.A. (1969). Percentage points of a test for clusters. Journal of the American Statistical Association, 64, 1647-1648.

[8] Everitt, B.S. (1979). Unsolved problems in cluster analysis. Biometrics, 35, 169-181.

[9] Friedman, H.P. and Rubin, J. (1967). On some invariant criterion for grouping. Journal of the American Statistical Association, 62, 1159-1178.

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

[11] Ganesalingam, S. and McLachlan, G.J. (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.

[12] Ganesalingam, S. and McLachlan, G.J. (1979). A case study of two clustering methods based on maximum likelihood. Statistica Neerlandica, 33, 81-90.

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

[14] Hartley, H.O. and Rao, J.N.K. (1968). Classification and estimation in analysis of variance problems. Review of International Statistical Institute, 36, 141-147.

[15] Hernandez-Alvi, A. (1979). Problems in Cluster Analysis. Unpublished D. Phil. thesis, University of Oxford.

[16] 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.

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

[18] Kiefer, N. (1978). Discrete parameter variation: efficient estimation of a switching regression model. Econometrika, 46, 427-434.

[19] Lee, K.L. (1979). Multivariate tests for clusters. Journal of the American Statistical Association, 74, 708-714.

[20] MacQueen, J. (1966). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

[21] Marriott, F.H.C. (1971). Practical problems in a method of cluster analysis. Biometrics, 27, 501-514.

[22] Marriott, F.H.C. (1974). The Interpretation of Multiple Observations. Academic Press, London.

[23] Marriott, F.H.C. (1975). Separating mixtures of normal distributions. Biometrics, 31, 767-769.

[24] McLachlan, G.J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation. Journal of the American Statistical Association, 70, 365-369.

[25] McLachlan, G.J. (1977). Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. Journal of the American Statistical Association, 72, 403-406.

[26] McLachlan, G.J. and Ganesalingam, S. (1980). Updating a discriminant function on the basis of unclassified data. Technical Report No. 47 , Department of Statistics, Stanford University.

[27] Murray, G.D. and Titterington, D.M. (1978). Estimation problem with data from a mixture. Applied Statistics, 27, 325-334.

[28] O'Neill, T.J. (1978). Normal discrimination with unclassified data. Journal of the American Statistical Association, 73, 821-826.

[29] Quandt, R.E. and Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions. Journal of the American Statistical Association, 73, 730-738.

[30] Sclove, S.L. (1977). Population mixture models and clustering algorithms. Communications in Statistics - Theory and Methods A6, 417-434.

[31] Scott, A.J. and Symons, M.L. (1971). Clustering methods based on likelihood ratio criteria. Biometrics, 27, 387-397.

[32] Symons, M.J. (1980). Clustering criteria for multivariate normal mixtures. Biometrics, 37 (to appear).

[33] Titterington, D.M. (1976). Updating a diagnostic system using unconfirmed cases. Applied Statistics, 25, 238-247.

[34] Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 5, 329-350.

[35] Wolfe, J.H. (1971). A Monte-Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Bulletin STB 72-2, Naval Personnel and Training Research Laboratory, San Diego.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 299	2. GOVT ACCESSION NO. AD-A203 084	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* THE CLASSIFICATION AND MIXTURE MAXIMUM LIKELIHOOD APPROACHES TO CLUSTER ANALYSIS		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) G. J. McLACHLAN		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS OFFICE Of Naval Research Statistics & Probability Program Code 436 Arlington, VA 22217		12. REPORT DATE MARCH 12, 1981 13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*)		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Cluster analysis, maximum likelihood approach, multivariate normal distributions.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

PLEASE SEE REVERSE SIDE.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

299

THE CLASSIFICATION AND MIXTURE MAXIMUM LIKELIHOOD APPROACHES TO CLUSTER ANALYSIS

A review is undertaken of two maximum likelihood approaches to cluster analysis, the so-called classification and mixture maximum likelihood methods. The basic assumptions of the two approaches and their associated properties are contrasted, in particular for multivariate normal component distributions. The problem of deciding how many clusters there are is discussed for each approach. Also, an account is given of the relative efficiency of the mixture approach to clustering.