INTRODUCTION TO THE USE OF MIXTURE MODELS IN CLUSTERING

K.E. BASFORD

BU-920-M December 1986

* Partially supported by Mathematical Sciences Institute and by the Australian-American Education Foundation.

ABSTRACT

The technique of clustering uses the measurements on a set of elements to identify clusters or groups of elements, such that there is relative homogeneity within the groups and heterogeneity between the groups. Under the mixture model approach, the elements are assumed to be a sample from a mixture of several populations in various proportions. This technique, in particular the case when the density function in each underlying population is assumed to be normal, is discussed in relation to other clustering techniques in common use.

It is suggested that this report be read in conjunction with the accompanying technical report "Illustrative examples of clustering using the mixture method and two comparable methods from SAS" by K.E. Basford, W.T. Federer and N.J. Miles-McDermott. There two real data sets are analysed using:

KMM	-	Normal mixture model method
SAS (CLUSTER)	-	Ward's method
SAS (CLUSTER)	-	EML method

and the results compared.

1. INTRODUCTION

A vast number of methods of clustering a set of elements into groups, such that there is relative homogeneity within the groups and heterogeneity between the groups, has been proposed. Recently, considerable emphasis has been placed on the use of mixture models where it is assumed that the elements have been sampled from a mixture of several populations in various proportions. This approach is considered here in the particular case where the underlying parametric form is the normal distribution.

In this report, the basis for the mixture model approach to clustering is discussed. A brief review of the general classification problem is given to place this particular technique in perspective. Then the formal definition of the mixture maximum likelihood method of clustering is given. The practical application is found in the accompanying technical report "Illustrative examples of clustering using the mixture method and two comparable methods from SAS" by K.E. Basford, W.T. Federer and N.J. Miles-McDermott. Much of this current report is to appear in a detailed study of mixture models in the book *Mixture Models: Inference and Applications to Clustering*, by G.J. McLachlan and K.E. Basford.

2. MIXTURE MODELS IN CLUSTERING

The technique of clustering uses the measurements on a set of elements to identify clusters or groups, in which the elements are relatively homogeneous, while they are heterogeneous between the clusters. The establishment of such clusters should enable a better perception and understanding of the information obtained on the elements, by observing the structure and relativities of these clusters. This method of analysis has been used in many scientific disciplines, including the biological sciences. There the situation is especially intricate because of the complex structure of the underlying biological mechanisms. Many interdependencies occur, and multidimensional measurement spaces are commonly encountered. Even if the elements being considered do not really consist of distinct groups, it still may be a useful and worthwhile exercise to cluster them into groups. A convenient labeling scheme may be all that is required, though usually, it is hoped that the particular grouping obtained may shed light on the phenomena of interest.

Vast numbers of clustering techniques have been proposed, and recently, considerable emphasis has been placed on the use of mixture models. Under this approach, it is assumed that the observations can be considered as a sample from a mixture of several populations in various proportions. Estimates of the distributions of the underlying populations (components) can then be obtained using the likelihood

principle, and the elements can be allocated to these populations on the basis of their estimated posterior probabilities. The mixture method is model based, in that the form of the density of an observation in each of the underlying populations has to be specified. A common approach is to take the component densities to be multivariate normal. The estimates of the parameters obtained may not be reliable if the sample is not large, nor, if there are departures from normality. However, some empirical studies (Hernandez-Avila, 1979) suggest that the mixture method applied with normal component densities may be fairly robust from the clustering view-point of being able to separate data in the presence of multimodality.

The history of the problem of decomposing a mixture is a long one, and there are many references concerned with mixtures of distributions (Gupta and Huang, 1981). The initial approach to this problem in the context of two univariate normal populations was considered by Karl Pearson (1894), who put forward a solution based on moments. Current thinking and experience have shown that other methods of estimation, most notably maximum likelihood (first used by Rao, 1948), are superior to the method of moments; see, for example, Tan and Chang (1972); Fryer and Robertson (1972); Holgersson and Jorner (1978). The maximum likelihood solution for a mixture of multivariate normal populations with a common covariance matrix was put forward by Day (1969). Wolfe (1970, 1971) studied mixtures of normal

distributions with unequal covariance matrices and mixtures of binomials. However, the parameter estimates cannot be obtained explicitly, and the convergence properties of the various iterative methods of solution were generally uncertain. It was not until Dempster, Laird and Rubin (1977) formalized this approach in a general context, through their EM algorithm, that the convergence properties were established on a theoretical basis.

Since then, several authors have utilized the mixture maximum likelihood approach for clustering purposes. Aitkin (1980) studied this technique for both parameter estimation and clustering in the two group context. Aitkin, Anderson and Hinde (1981) presented a detailed statistical modeling of an extensive body of research data on teaching styles, in which they clustered teachers into groups. They felt mixture models were an appropriate and useful tool, as "when clustering samples from a population, no cluster method is *a priori* believable without a statistical model". Also, as they pointed out, "cluster methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory".

Before proceeding with the formal definition of the mixture maximum likelihood method of clustering, a brief review of the general classification problem is given to place this particular technique in perspective.

3. BACKGROUND TO THE GENERAL CLASSIFICATION PROBLEM

Firstly, it is important to establish a standard terminology to describe the data sets which will be considered. Carroll and Arabie (1980) introduced "a taxonomy of measurement data", in which, a mode is defined as a particular class of entities, and an N-way array is defined as the cartesian product of a number of modes, some of which may be repeated. Thus, if the data consist of the measurements of certain characteristics of the elements, then the appropriate description is two-mode two-way data; one mode being the elements and the other being the characteristics. If, however, the data are in the form of proximities between the elements, based on the above measurements, then it would be described as one-mode two-way data; the one mode being the elements. In both cases the data would be displayed in a two-way table, that is, rows by columns. The former is a more informative type of basic data set as it can be easily converted, if required, to the latter, by suitable definition of a similarity or dissimilarity measure.

Consider such a two-mode two-way array, where p attributes have been measured on each of n elements. The problem is to classify these elements into g groups, such that the elements within a group are, in some sense, homogeneous. If existence of the groups is known, and there are available data of known origin from each of the groups for constructing estimates of the group densities, then a sample based allocation rule can be formed for assigning the

elements of unknown origin to the possible groups with the minimum probability of misclassification. This discriminant analysis problem has been well studied, and the reader is directed to Kshirsagar (1972), Lachenbruch (1975) and Lachenbruch and Goldstein (1979), and the references within. In contrast to this, cluster analysis is the multivariate technique used to create groups amongst the elements, where there is no prior information regarding the underlying group structure, or at least, where there are no available data from each of the groups if their existence is known.

The need for cluster analysis has arisen in a natural way in many fields of study. In the last twenty years, the quantity of literature on this topic has grown enormously, but unfortunately it has been mainly intra-disciplinary. This lack of inter-disciplinary communication has meant that large bodies of researchers appear to be unaware of one another (Anderberg, 1973). Noteworthy attempts at classifying and reviewing cluster methods appear in Cormack (1971), Das Gupta (1973), Anderberg (1973), Sneath and Sokal (1973), Everitt (1978, 1979, 1980) and Mezzich and Solomon (1980), while various appoaches to cluster analysis are considered in van Ryzin (1977).

Most clustering techniques are appropriate to data that are in the form of a two-mode two-way array (p measurements on each of n elements), or a one-mode two-way array (proximities measured between n elements), as described earlier. Also, they assume that the initial sample is

unstructured, in the sense, that there are no replications of any particular element specifically identified as such, and that all elements are independent of one another. Within this framework, available methods of seeking clusters can be categorized broadly as being hierarchical or non-hierarchical. The former class is one in which every cluster obtained at any stage is a merger or split of clusters at other stages. Thus, it is possible to visualize not only the two extremes of clustering, that is n clusters with one element per cluster (weak clustering) and a single cluster with all n elements (strong clustering), but also a monotonically increasing strength of clustering as one goes from one level to another. A hierarchical strategy always optimizes a route between these two extremes (Williams, 1976). The route may be defined by progressive fusions, beginning with n single element groups and ending with a single group of n elements (agglomerative hierarchy); or by progressive divisions, beginning with a single group and decomposing it into individual elements (divisive hierarchy).

Agglomerative hierarchical clustering has been studied by Ward (1963), Sokal and Sneath (1963), Hartigan (1967), Johnson (1967), and Wishart (1968, 1969), among many others. There have been numerous investigations of the applicability of various agglomerative hierarchical techniques to simulated data with differing properties. Kuiper and Fisher (1975) and Mojena (1977) both recommended Ward's minimum variance method. Milligan and Isaac (1980) felt these investigations

were not generally valid because diagonal covariance matrices were used in generating the data. They performed an extensive simulation study to compare such methods, and found that Ward's method did not perform as well as some other algorithms, for example single linkage (nearest neighbour). Bayne et al. (1980) used non-diagonal covariance matrices, and came to the conclusion that non-hierarchical methods were only slightly better than some of the hierarchical techniques, in particular Ward's method.

Williams (1976) noted that all agglomerative strategies suffer from two disadvantages, the first of which is computational. The user's interest is normally concentrated in the higher levels of the hierarchy, so that it is almost invariably necessary to establish the complete hierarchy from individual elements to a single group of all elements. Secondly, an agglomerative system is inherently prone to a small amount of misclassification, the ultimate cause of which is that the process begins at the inter-individual level, where the possibility of this type of error is greatest. Divisive classifications (Edwards and Cavalli-Sforza, 1965) are free of these disadvantages, but are not straight forward to apply, save in the case of a monothetic system when a single attribute is used to cluster the elements (Williams, 1976). Carmer and Lin (1983) compared five univariate divisive clustering methods for grouping means in analysis of variance, and found them to be particularly dependent on the precision of the experiment,

rather than the stated significance level or clustering method used. In contrast, a polythetic system is one based on a measure of similarity or dissimilarity applied over all observed attributes, so that an element is grouped with those elements which, on the average, it most resembles.

In non-hierarchical procedures, new clusters are obtained by both lumping and splitting of old clusters, and although the two extremes of clustering are still the same, the intermediary stages of clustering do not have the natural monotone character of strength of clustering. Thus with a non-hierarchical strategy, it is the structure of the individual groups which is optimized, since these are made as homogeneous as possible (Williams, 1976). No route is defined between the groups and their consitituent elements, so that the infrastructure of a group cannot be examined in this way. For those applications in which homogeneity of groups is of prime importance, the non-hierarchical strategies are very attractive. Marriott (1974, 1982), Gnanadesikan (1977) and Everitt (1978) have given excellent discussions of these procedures. A crucial question here is the computational feasibility of any specific algorithm. An examination of all possible partitions of the data, to determine a clustering or grouping that is optimal with respect to some criterion, is prohibitively expensive, and may be impossible despite the speed of today's computers (Gnanadesikan, 1977).

To illustrate criteria used in non-hierarchical cluster techniques, let T be the total scatter matrix initially

defined by Wilks (1962). Then for each partition of n elements into g groups, T can be expressed as the sum of W, the pooled within group scatter matrix, and B, the between group scatter matrix. For a given set of elements, T is fixed, so a natural criterion for grouping is to minimize W or equivalently maximize B (Edwards and Cavalli-Sforza, 1965) and Singleton (1965, as reported by Friedman and Rubin, 1967) achieved this by minimizing trace W. MacQueen (1967) and Hartigan (1975, 1978) used the so-called k-means procedure which is a special implementation of the trace W criterion. As mentioned earlier, many clustering procedures start with an nxn symmetric matrix of pairwise distances or similarities between elements. If the trace W criterion is chosen, then so implicitly is ordinary Euclidean distance, as trace W can be computed directly from these pairwise distances (Friedman and Rubin, 1967). Wilks (1962) introduced $|W|/|T|$ as a statistic, and Friedman and Rubin (1967) maximized its reciprocal $|T|/|W|$. Another related criterion function is the maximum of trace $(W^{-1}B)$. This is sometimes called Hotelling's Trace Criterion, and is equivalent to what Rao (1952) called the generalization to $g>2$ groups of the Mahalanobis distance between two groups. As stated by Friedman and Rubin (1967), both trace $(W^{-1}B)$ and $|T|/|W|$ may be expressed in terms of the eigenvalues of $W^{-1}B$, and Anderson (1958) showed that these eigenvalues are the only invariants of W and B under non-singular linear transformations of the original data matrix. While trace W

is only invariant under an orthogonal transformation, $|T|/|W|$ is invariant under any non-singular transformation (Friedman and Rubin, 1967). Also, the trace W criterion does not take into account the within group covariance structure of the measurements, and though computationaly simpler, is less likely to identify elongated clusters than the $|W|$ criterion (Marriott, 1971). In addition, Friedman and Rubin (1967) found that the latter criterion demonstrated greater sensitivity to the local structure of data considered in their investigations.

Scott and Symons (1971) showed that these common non-hierarchical clustering procedures were extensions of the likelihood ratio method of classification for normal populations, where the unknown indicator variables associated with the data are treated as unknown parameters to be estimated along with the other unknown parameters by maximum likelihood. In particular, for known equal spherical covariance matrices, the maximum likelihood partition corresponds to minimizing trace W, while for unknown equal, but not necessarily spherical covariance matrices, the maximum likelihood partition is equivalent to minimizing $|W|$. Symons (1981) discussed, in some detail, such criteria derived from maximum likelihood and Bayesian approaches corresponding to different assumptions about the covariance matrices of the underlying component populations.

Hawkins, Muller and ten Krooden (1982, page 353) commented that most writers on cluster analysis "lay more

stress on algorithms and criteria in the belief that intuitively reasonable criteria should produce good results over a wide range of possible (and generally unstated) models". For example, the trace W criterion is predicated on normal data with spherical within-cluster covariance matrices as noted above, but as they pointed out, many users would apply this criterion even in the face of contrary evidence. They strongly supported the increasing emphasis on a model based approach to clustering. Mixture models have thus been the subject of recent attention for use in this context. In particular, the mixture maximum likelihood method provides a concise way of summarizing differences among the elements being considered. It is therefore worthwhile considering this approach, particularly in situations where there is some doubt about the validity of the clusters obtained by some other method (see Aitkin, Anderson and Hinde (1981) and the subsequent discussion, and Aitkin (1983) on the role of the mixture approach versus less complicated methods based on mean analysis).

With the mixture maximum likelihood approach, it is assumed that a p-dimensional observation is available for each of n elements, assumed to have been drawn from a mixture of a specified number of populations (groups) in various proportions. By adopting some parametric form for the density function in each underlying population, a likelihood can be formed in terms of the mixture density, and the unknown parameters estimated by the likelihood principle. An

allocation rule based on the estimated posterior probabilities can then be formed for assigning the elements to their unknown population of origin. The properties of the mixture approach have been considered by Day (1969), Wolfe (1970), Hosmer (1973a), O'Neill (1978), Ganesalingam and Mclachlan (1978, 1980a, 1980b, 1981), Aitkin (1980), Mezzich and Solomon (1980), Aitkin, Anderson and Hinde (1981), Symons (1981), and Everitt and Hand (1981), among many others. In particular, Ganesalingam (1980) studied the mixture maximum likelihood approach to estimation and clustering in the two group context in a Ph.D. dissertation at the University of Queensland. Much of this and the associated work were essentially summarized by McLachlan (1982). More recently, Basford (1985) investigated cluster analysis via normal mixture models in the more general case of an unrestricted number of groups.

The general problem of validating clustering results has become of increasing importance (Dubes and Jain, 1979; Murtagh, 1983), regardless of which clustering technique is employed. This is particularly difficult as, in cluster analysis, the origin of each element is unknown. Based on ideas developed in the discriminant analysis context, Ganesalingam (1980) showed that in the case of g=2, estimates of error rates can be obtained to assess the overall performance of the mixture maximum likelihood method of clustering. Such estimates are based on the maximum of the estimated posterior probabilities of the elements belonging

to the various populations. This facility for assessing performance is highly desirable, and is further developed in the cluster analysis context in Basford and McLachlan (1985a).

The mixture approach has the potential to handle structured data because it is model based. The structure being referred to here is with respect to the collection and presentation of the data before a clustering technique is to be applied. It is not with reference to the underlying structure among the elements which the clustering technique is being used to identify. The structure of the data could be in the form of repeated observations on each element by observing them in some experimental design, or it could be the representation of the information on the elements in the three-way array. Most clustering techniques assume the data are in the form of a one-mode or two-mode two-way array with no repeated observations as such. Hence the data have to be reduced to this form before a clustering technique can be applied. To illustrate these points, consider how clustering methods are currently utilized in two relevant examples of biological data.

In the first example, suppose a large number of treatments of some description are being compared in an experimental design suitable for analysis of variance. The researcher may decide that it would be useful, and perhaps even sufficient, to split these treatments into relatively homogeneous groups, rather than to compare each individual

treatment. Thus, here it is the treatments that are being considered as elements which are to be clustered into groups on the basis of a univariate attribute. A common approach is to reduce the observed data to information on the mean for each treatment before a cluster technique is applied. Scott and Knott (1974) and Carmer and Lin (1983) used hierarchical techniques to cluster such treatment means. Binder (1978, 1981) and Menzefricke (1981) adopted a Bayesian approach. Skillings (1983) considered a non-parametric approach to comparing means in a one-way analysis of variance, while Cox and Spjotvoll (1982) devised a method of partitioning means into groups based on standard F tests. Aitkin (1980) showed how the mixture method could be used to cluster treatment means from a one-way experimental design via the EM algorithm of Dempster, Laird and Rubin (1977). Because this is a model based technique, it can be used to analyze the data without necessarily reducing it to means (Basford and McLachlan, 1985b). This could be relevant when more complicated statistical designs with non-independent observations have been employed. In this example only univariate data have been considered, but there appears no reason why multivariate data could not be considered similarly.

The second example concerns data sets which are in the form of three-mode three-way arrays. Consider the results of a large plant improvement program expressed as a genotype by attribute by environment matrix (Basford, 1982). This is quite typical of experiments where various attributes are

measured on each of a large number of genotypes grown in several environments. The aim of the cluster analysis is to obtain a suitable grouping of the genotypes, as a convenient labeling scheme, and to shed light on the underlying relationships between the genotypes. As stated earlier, most clustering techniques require the data in the form of a two-way array; a genotype by attribute array, obtained by averaging over environments, or else, a genotype by environment array for each attribute may be used. In the latter case, a cluster analysis would have to be performed for each attribute of interest. Examples of such analyses are given by Burt et al. (1971), Mungomery, Shorter and Byth (1974) and Byth, Eisemann and DeLacy (1976). If, however, all the information collected was pertinent to the clustering of the genotypes, then it would seem to be an advantage if a clustering technique could handle the entire three-way array in a single analysis.

It may be possible to combine attributes to produce a single measure which would then enable the data to be represented by a two-mode two-way array. For example, a new variable, energy yield, might be defined as the addition of protein percentage and oil percentage, each multiplied by seed yield. Another example would be the use of selection indices (Smith, 1936; Manning, 1956). However, a suitable combination of attributes connot always be determined, and it is then more appropriate to consider the attributes as individually contributing information to the formation of the

clusters. Similarly, variable reduction techniques, which convert the data to a two-way array, appear to be circumventing the problem of determining a method of clustering to analyze data directly in the form of a three-mode three-way array.

Because of the perceived inability of methods of cluster analysis to handle three-way arrays, some researchers have turned to the technique of multidimensional scaling (MDS) to obtain a low dimensional spatial representation (Torgerson, 1958). It has been widely used in the social and behavioral sciences as a descriptive model for elucidating data patterns (Kruskal and Wish, 1978), and was extended to cover three-way tables of the type described above (Tucker and Messick, 1963; Carroll and Chang, 1970). Using the individual differences model of Carroll and Chang (1970), Basford (1982) analyzed soybean data by postulating that an underlying pattern of genotype performance, as measured by an array of attributes across environments, existed, and that there was an underlying space of small dimension, in which the genotypes could be placed. Under this model, the position of the genotypes, as determined by the environments, may vary only because of change in the relative importance of these conceptual underlying axes. The relative position of the points (genotypes here) in this space was then used as an indication of similarity of response pattern. The MDS approach is not attempting to place the elements into discrete groups, but rather to obtain a low dimensional

spatial representation. It, therefore, is not a competing technique, but rather a complementary one to clustering (Kruskal, 1977).

Recently, there have been some new developments in clustering techniques, which attempt to use the individual differences concept, as introduced in MDS by Carroll and Chang (1970), to enable the processing of three-way data. Carroll and Arabie (1983) devised a method for non-hierarchical overlapping clustering called INDCLUS, in which each of a number of subjects or individual data sources perceive a common set of clusters of elements, but these clusters are differentially weighted by subjects in order to portray individual differences. Carroll, Clark and DeSarbo (1984) developed a new methodology called INDTREES for a hierarchical tree structure to obtain a discrete network representation of such three-way data. In their model, the individual differences generalization is one in which subjects or individual data sources are assumed to base their judgements on the same family of trees, but are allowed to have different node heights and/or branch lengths.

Basford and McLachlan (1985c) appear to have been the first to consider the mixture method of clustering in relation to data in the form of three-mode three-way arrays. As it is a model based technique, this approach to clustering does have the ability to handle such structure. In particular, the genotype by environment interaction, which is of considerable importance in large plant breeding trials,

can be directly incorporated into this model, as shown in the above paper.

4. GENERAL DEFINITION OF THE MIXTURE MAXIMUM LIKELIHOOD APPROACH

Multivariate observations on a set of n elements forming a two-mode two-way array can be represented as $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$. In applying the mixture method of clustering, it is assumed in the first instance that there is a specified number, say g, of underlying populations $\Pi_1,\ldots,\Pi_g$. It is then assumed that the sample $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$ has been drawn from the superpopulation Π, a mixture of these underlying populations. The proportions in which the populations are represented in the mixture are unknown, and will be denoted by $\underset{\sim}{\pi}=(\pi_1,\ldots,\pi_g)'$. Let the density of an observation $\underset{\sim}{x}$ from Π_i be given by $f_i(\underset{\sim}{x};\ \underset{\sim}{\upsilon})$ where $\underset{\sim}{\upsilon}$ denotes the vector of unknown parameters. The mixture method of clustering can be applied, at least in principle, provided the form of these densities is known. The most widely studied examples of this formulation concern random samples from a mixture of normal distributions; see Rao (1948), Hill (1963), Hasselblad (1966), Choi (1969a, 1969b), Day (1969), Wolfe (1970, 1971), Urbakh (1972), Dick and Bowden (1973), Hosmer (1973a, 1973b, 1974) and Kazakos (1977). Hasselblad (1969) treated more general random sampling models, giving as examples mixtures of Poissons, binomials, and exponentials. Symons, Grimson and Yuan (1983) considered a mixture of Poisson

distributions, while Aitkin (1980) clustered multinomial observations. The special issue of Communications in Statistics on remote sensing, published in 1976, gives additional references, especially with regard to estimating mixing proportions. A summary of the work contained in most of these is given by James (1978).

An observation $\underset{\sim}{x}$ in Π has the mixture density given by

$$f(\underset{\sim}{x};\ \underset{\sim}{v},\underset{\sim}{\pi}) = \sum_{i=1}^{g} \pi_i f_i(\underset{\sim}{x};\ \underset{\sim}{v}). \tag{4.1}$$

Anderson (1972) called this a compound distribution, so as to avoid the confusion that can arise in using the word mixture in the context of mixture sampling. The likelihood of the n observations is given by

$$L = \prod_{j=1}^{n} \{ \sum_{i=1}^{g} \pi_i f_i(\underset{\sim}{x}_j;\ \underset{\sim}{v}) \}. \tag{4.2}$$

The vector $\underset{\sim}{\phi}'=(\underset{\sim}{\Pi}',\underset{\sim}{v}')'$ of unknown parameters can be estimated using the likelihood principle. Then each $\underset{\sim}{x}_j$ can be allocated on the basis of its estimated posterior probabilities of belonging to the various populations. The posterior probability that $\underset{\sim}{x}_j$, (really the element with observation $\underset{\sim}{x}_j$), belongs to Π_i is given by

$$\theta_i(\underset{\sim}{x}_j;\ \underset{\sim}{\varphi}) = \pi_i f_i(\underset{\sim}{x}_j;\ \underset{\sim}{v}) \Big/ \sum_{u=1}^{g} \pi_u f_u(\underset{\sim}{x}_j;\ \underset{\sim}{v}), \quad (i = 1,\ldots,g). \tag{4.3}$$

It is estimated by replacing the unknown parameter vector $\underset{\sim}{\varphi}$ with the likelihood estimate $\hat{\underset{\sim}{\varphi}}$. Then $\underset{\sim}{x}_j$ is assigned to Π_u if

$$\theta_u(\underset{\sim}{x}_j;\ \hat{\underset{\sim}{\varphi}}) > \theta_i(\underset{\sim}{x}_j;\ \hat{\underset{\sim}{\varphi}}), \ (i = 1,\ldots,g;\ \ i \neq u). \tag{4.4}$$

For convenience, $\theta_i(\underset{\sim}{x}_j;\ \hat{\underset{\sim}{\varphi}})$ is denoted by $\hat{\theta}_{ij}$ while $\theta_i(\underset{\sim}{x}_j;\ \underset{\sim}{\varphi})$ is denoted by θ_{ij} . If $\underset{\sim}{\varphi}$ was known, the allocaton rule (4.4) would be the optimal or Bayes rule (Anderson, 1958) which minimizes the overall error rate.

The likelihood equation for $\underset{\sim}{\varphi}$, $\delta \log L/\delta \underset{\sim}{\varphi} = \underset{\sim}{0}$, can be expressed as

$$\sum_{i=1}^{g} \sum_{j=1}^{n} \hat{\theta}_{ij}\ \delta \log f_i(\underset{\sim}{x}_j;\ \underset{\sim}{v})/\delta \hat{\underset{\sim}{v}} = \underset{\sim}{0} \tag{4.5}$$

and

$$\hat{\pi}_i = \sum_{j=1}^{n} \hat{\theta}_{ij}/n, \qquad (i = 1,\ldots,g). \tag{4.6}$$

When the maximum likelihood estimates exist, the computation is facilitated by identifying these equations with the application of the EM algorithm of Dempster, Laird and Rubin (1977). They discussed this problem in a very general context where the populations are mixed with respect to a distribution whose parameters may be related to the population parameters. In the current model, the mixing proportions, $\pi_1,.....\pi_g$, are unrelated to the population parameters in $\underset{\sim}{\upsilon}$. For each $\underset{\sim}{x}_j$, let the vector of indicator variables, $\underset{\sim}{\tau}_j = (\tau_{1j},...,\tau_{gj})'$, be defined by

$$\tau_j = \begin{cases} 1, & \underset{\sim}{x}_j \in \Pi_i \\ 0, & \underset{\sim}{x}_j \notin \Pi_i \end{cases} . \qquad (4.7)$$

The expectation of τ_{ij} conditional on $\underset{\sim}{x}_j$ is equal to θ_{ij}. Then it can be verified that equations (4.5) and (4.6) are obtained by differentiation of the expectation of the complete data log likelihood conditional on $\underset{\sim}{x}_1,...,\underset{\sim}{x}_n$. This conditional expectation is effected here by replacing each indicator variable τ_{ij} by its expectation conditional on $\underset{\sim}{x}_j$; that is, θ_{ij} .

The iterative process follows in two steps. First (the E step), given some initial value for the vector of estimates, say $\underset{\sim}{\varphi}^{(o)}$, the τ_{ij} are estimated by

$$E(\tau_{ij}|\underset{\sim}{x}_j;\ \varphi^{(o)}) = Pr(\underset{\sim}{x}_j \in \Pi_i|x_j;\ \varphi^{(o)})$$

$$= \theta_i(\underset{\sim}{x}_j;\ \varphi^{(o)}), \quad (i = 1,\ldots,g). \qquad (4.8)$$

Second (the M step), for the estimated τ_{ij} , φ , say $\varphi^{(1)}$, is chosen to maximize the likelihood. The E and M steps are alternated repeatedly to give a sequence $\{\varphi^{(q)}\}$. It follows that

$$L(\varphi^{(q+1)}) \geq L(\varphi^{(q)}). \qquad (4.9)$$

and so if bounded above, $L(\varphi^{(q)})$ converges to some L^* which will be a local maximum, provided the sequence is not trapped at some saddle point (Wu, 1983; Boyles, 1983). Generally the convergence is slow, but may be improved using Aitken's acceleration process; see Louis (1982) for details of speeding up this algorithm. With mixture models, the likelihood often has multiple maxima, and so the EM algorithm should be repeated for several different sets of starting values of φ. In McLachlan and Basford (1987) there is a discussion on the choice of suitable starting values during the search for all local maxima, and on the problem of which of these to choose.

With the solution of the likelihood equation under the mixture approach, there is no insistence on outright allocation of the elements to the groups at each stage of the iterative process, thus avoiding the inconsistent estimates as obtained with, say, the $|W|$ criterion. Providing regularity conditions hold, the estimates so obtained have the desirable large sample properties of likelihood estimators; for example, consistency, asymptotic efficiency and normality.

REFERENCES

Aitkin, M. (1980). Mixture applications of the EM algorithm in GLIM. *Comstat 1980: Proceedings in Computational Statistics*, 537-541. Vienna: Physica-Verlag.

Aitkin, M. (1983). Comment on paper by S.J. Prais. *Journal of the Royal Statistical Society* A **146**, 170-171.

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society* A **144**, 419-461.

Anderberg, M.R.C. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **16**, 31-50.

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons.

Basford, K.E. (1982). The use of multidimensional scaling in analysing multi-attribute genotype response across environments. *Australian Journal of Agricultural Research* **33**, 473-480.

Basford, K.E. (1985). *Cluster analysis via normal mixture models*. Unpublished Ph.D. thesis, University of Queensland.

Basford, K.E., Federer, W.T. and Miles-McDermott, N.J. (1987). Illustrative examples of clustering using the mixture method and two comparable methods from SAS.

Cornell University Biometrics Unit Technical Report BU-921-M, Ithaca, New York.

Basford, K.E. and McLachlan, G.J. (1985a). Estimation of allocation rates in a cluster analysis context. *Journal of the American Statistical Association* 80, 286-293.

Basford, K.E. and McLachlan, G.J. (1985b). Cluster analysis in a randomized complete block design. *Communications in Statistics - Theory and Methods* 14, 451-463.

Basford, K.E. and McLachlan, G.J. (1985c). The mixture method of clustering applied to three-way data. *Journal of Classification* 2, 109-125.

Bayne, C.K., Beauchamp, J.J., Begovich, C.L. and Kane, V.E. (1980). Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition* 12, 51-62.

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

Binder, D.A. (1981). Approximations to Bayesian clustering rules. *Biometrika* 68, 275-285.

Boyles, R.A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society B* 45, 47-50.

Burt , R.L., Edye, L.A., Williams, W.T., Grof, B. and Nicholson , C. H. L. (1971). Numerical analysis of variation patterns in the genus *Stylosanthes* as an aid to plant introduction and assessment. *Australian Journal of Agricultural Research* 22, 737-757.

Byth, D.E., Eisemann, R.L. and DeLacey, I.H. (1976). Two-way pattern analysis of a large data set to evaluate genotypic adaptation. *Heredity* 37, 215-230.

Carmer, S.G. and Lin, W.T. (1983). Type I error rates for divisive clustering methods for grouping means in analysis of variance. *Communications in Statistics - Simulation and Computation* 12, 451-466.

Carroll , J.D. and Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology* 31, 607-649.

Carroll, J.D. and Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika* 48, 157-169.

Carroll , J.D. , and Chang , J.J. (1970). Analysis of individual differences in multidimensional scaling via an *N*-way generalization of Eckart-Young decomposition. *Psychometrika* 35, 283-319.

Carroll, J.D., Clark, L.A. and DeSarbo, W.S. (1984). The representation of three-way proximity data by single and multiple tree structure models. *Journal of Classification* 1, 25-74.

Choi, K. (1969a). Estimators for the parameters of a finite mixture of distributions. *Annals of the Institute of Statistical Mathematics* 21, 107-116.

Choi, K. (1969b). Empirical Bayes procedure for (pattern) classification with stochastic learning. *Annals of the Institute of Statistical Mathematics* 21, 117-125.

Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society* A **134**, 321-367.

Cox, D.R. and Spjotvoll, E. (1982). On partitioning means into groups. *Scandanavian Journal of Statistics* **9**, 147-152.

Das Gupta, S. (1973). Theories and methods in classification: A review. In *Discriminant Analysis and Application*. Ed. T. Cacoullos, New York: Academic Press, 77-138.

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

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 B* **39**, 1-38.

Dick, N.P. and Bowden, D.C. (1973). Maximum likelihood estimation for mixtures of two normal distributions. *Biometrics* **29**, 781-790.

Dubes, R. and Jain, A.K. (1979). Validity studies in clustering methodologies. *Pattern Recognition* **11**, 235-254.

Edwards, A.W.F. and Cavalli-Sforza, L.L. (1965). A method for cluster analysis. *Biometrics* **21**, 362-375.

Everitt, B.S. (1978). *Graphical Techniques for Multivariate Data*. London: Heinemann Educational Books Ltd.

Everitt , B.S. (1979). Unresolved problems in cluster analysis. *Biometrics* **35**, 169-181.

Everitt, B.S. (1980). *Cluster Analysis*. 2nd Edition. London: Wiley-Halsted.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

Friedman, H.P. and Rubin, J. (1967). On some invariate criteria for grouping data. *Journal of the American Statistical Association* 62, 1159-1178.

Fryer, J.G. and Robertson, C.A. (1972). A comparison of some methods for estimating mixed normal distributions. *Biometrika* 59, 639-648.

Ganesalingam, S. (1980). *On the mixture maximum likelihood approach to estimation and clustering*. Unpublished Ph.D. Thesis, University of Queensland.

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

Ganesalingam, S. and McLachlan, G.J. (1980a). A comparison of the mixture and classification approaches to cluster analysis. *Communication in Statistics - Theory and Methods* A 9, 923-933.

Ganesalingam, S. and McLachlan, G.J. (1980b). Error rate estimation on the basis of posterior probabilities. *Pattern Recognition* 12, 405-413.

Ganesalingam, S. and McLachlan, G.J. (1981). Some efficiency results for the estimation of the mixing proportion in a

mixture of two normal distributions. *Biometrics* 37, 23-33.

Gnanadesikan , R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: John Wiley and Sons.

Gupta , S.S. and Huang , W.T. (1981). On mixtures of distributions: A survey and some new results on ranking and selection. *Sankhyā*: *B* 43, 245-290.

Hartigan, J.A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association* 62, 1140-1158.

Hartigan, J.A. (1975). *Clustering Algorithms*. New York: John Wiley and Sons.

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

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics* 8, 431-444.

Hasselblad , V. (1969). Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association* 64, 1459-1471.

Hawkins, D.M. , Muller, M.W. and ten Krooden, J.A. (1982). Cluster analysis. In *Topics in Applied Multivariate Analysis*. Ed. D. M. Hawkins, Cambridge: Cambridge University Press, 303-356.

Hernandez-Avila, A. (1979). *Problems in Cluster Analysis.* Unpublished D. Phil. Thesis, University of Oxford.

Hill , B. M. (1963). Information for estimating the proportions in mixtures of exponential and normal distributions. *Journal of the American Statistical Association* 58, 918-932.

Holgersson, M. and Jorner, U. (1978). Decomposition of a mixture into normal components: A review. *International Journal of Bio-Medical Computing* 9, 367-392.

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

Hosmer, D.W. (1973b). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics* 29, 761-770.

Hosmer, D.W. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines. *Communications in Statistics* 3, 995-1006.

James, I.R. (1978). Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements. *Biometrics* 34, 265-275.

Johnson , S. C. (1967). Hierarchical clustering schemes. *Psychometrika* 32, 241-254.

Kazakos , D. (1977). Recursive estimation of prior probabilities using a mixture. *IEEE Transactions on Information Theory* IT-23, 203-211.

Kruskal , J.B. (1977). The relationship between multi-dimensional scaling and clustering. In *Classification and Clustering*. Ed. J. van Ryzin, New York: Academic Press, 17-44.

Kruskal, J.B. and Wish, M. (1978). *Multidimensional Scaling*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverley Hills: Sage Publications.

Kshirsager, A.M. (1972). *Multivariate Analysis*. New York: Marcel Dekker.

Kuiper, F.K. and Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics* 31, 777-783.

Lachenbruch, P.A. (1975). *Discriminant Analysis*. New York: Hafner Press.

Lachenbruch, P.A. and Goldstein, M. (1979). Discriminant Analysis. *Biometrics* 35, 69-85.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B* 44, 226-233.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics*

and Probability Vol. 1. Berkeley : University of California Press, 281-297.

Manning, H.L. (1956). Yield improvement from a selection index technique with cotton. *Heredity* 10, 303-322.

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

Marriott, F.H.C. (1974). *The Interpretation of Multivariate Observations*. New York: Academic Press.

Marriott, F.H.C. (1982). Optimization methods of cluster analysis. *Biometrika* 69, 417-421.

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

McLachlan, G.J. and Basford, K.E. (1987). *Mixture Models: Inference and Applications to Clustering*. To be published in New York by Marcel Dekker.

Menzefricke, U. (1981). Bayesian clustering of data sets. *Communications in Statistics - Theory and Methods* A 10, 65-77.

Mezzich , J. E. and Solomon , H. (1980). *Taxonomy and Behavioral Science - Comparative Performance of Grouping Methods*. New York: Academic Press.

Milligan, G.W. and Isaac, P.D. (1980). The validation of four ultrametric clustering algorithms. *Pattern Recognition* 12, 41-50.

Mojena , R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal* 20, 359-363.

Mungomery, V.E., Shorter, R. and Byth, D.E. (1974). Genotype x environment interactions and environmental adaptation. I. Pattern analysis - application to soya bean populations. *Australian Journal of Agricultural Research* 25, 59-72.

Murtagh, F. (1983). A probability theory of hierarchical clustering using random dendograms. *Journal of Statistical Computation and Simulation* 18, 145-157.

O'Neill, T.J. (1978). Normal discriminant with unclassified observations. *Journal of the American Statistical Association* 73, 821-826.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society* A 185, 71-110.

Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society B* 10, 159-203.

Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: John Wiley and Sons.

Scott, A.J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* 30, 507-512.

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

Skillings, J.H. (1983). Nonparametric approaches to testing and multiple comparisons in a one - way anova. *Communications in Statistics - Simulation and Computation* 12, 373-387.

Smith , H.F. (1936). A discriminant function for plant selection. *Annals of Eugenics* 7, 240-250.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy; the Principles and Practice of Numerical Classification.* San Francisco: W.H. Freeman.

Sokal, R.R. and Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*. San Francisco: W.H. Freeman.

Symons, M.J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics* 37, 35-43.

Symons , M.J. , Grimson , R.C. and Yuan , Y.C. (1983). Clustering of rare events. *Biometrics* 39, 193-205.

Tan, W.Y. and Chang, W.C. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association* 67, 702-708.

Torgerson, W.S. (1958). *Theory and Methods of Scaling*. New York: John Wiley and Sons.

Tucker , L.R. and Messick , S. (1963). An individual differences model for multidimensional scaling. *Psychometrika* 28, 333-367.

Urbakh, V.Yu. (1972). A discriminant method of clustering. *Journal of Multivariate Analysis* 2, 249-260.

Van Ryzin, J. (1977). *Classification and Clustering*. New York: Academic Press.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58, 236-244.

Wilks, S.S. (1962). *Mathematical Statistics*. New York: John Wiley and Sons.

Williams, W.T. (1976). Types of classification. In *Pattern Analysis in Agricultural Science*. Ed. W.T. Williams, Amsterdam: Elsevier Scientific Publishing Company, 76-83.

Wishart, D. (1968). A *Fortran II Program for Numerical Classification*. St. Andrews: St. Andrews University.

Wishart , D. (1969). An algorithm for hierarchical classification. *Biometrics* 25, 165-170.

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

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

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95-103.