An Algorithm for Unsupervised Learning via Normal Mixture Models

G.J. McLachlan* gjm@maths.uq.edu.au
D. Peel dp@maths.uq.edu.au
Department of Mathematics
The University of Queensland
Queensland 4072, Australia

ph: +07-3-365-2150
fax: +07-3-365-1477

Abstract: *We consider the approach to unsupervised learning whereby a normal mixture model is fitted to the data by maximum likelihood. An algorithm called NMM is presented that enables the normal mixture model with either restricted or unrestricted component covariance matrices to be fitted to a given data set. The algorithm automatically handles the problem of the specification of initial values for the parameters in the iterative fitting of the model within the framework of the EM algorithm. The algorithm also has the provision to carry a test for the number of components on the basis of the likelihood ratio statistic.*

Keywords: Mixture models, Maximum likelihood, EM algorithm, Likelihood ratio test.

Area of Interest: Concept Formation and Classification.

1 Introduction

In this paper we consider the development of an algorithm for the fitting of a normal mixture model in the absence of data on entities that have been classified with respect to the components of the mixture. This is usually referred to as unsupervised learning in the pattern recognition literature. By fitting a mixture model with g components to the data, we obtain a probabilistic clustering of the unclassified entities in terms of their estimated posterior probabilities of component membership. A partition of the entities into g nonoverlapping clusters is given by assigning an entity to the component to which it has the highest estimated (posterior) probability of belonging.

Mixture models can be fitted via maximum likelihood. This so-called mixture likelihood-based approach is about the only clustering technique that is satisfactory from the mathematical point of view. It assumes a well-defined model that can be fitted by the well-established statistical method of maximum likelihood. It also allows the important question of how many clusters there are in the data to be posed within the framework of standard statistical theory. Further, in a discriminant analysis context where there is *a priori* knowledge of the underlying group structure, it can be used to update an existing discriminant function on the basis of the unclassified data subsequently available (McLachlan (1992, Section 2.7)).

2 Mathematical Formulation

Suppose that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ denote the available data, where $\boldsymbol{y}_j$ denotes the observed value on some p-dimensional feature vector $\boldsymbol{Y}$ measured on the jth entity ($j = 1, \ldots, n$). With the mixture likelihood approach to clustering, the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are assumed to have come from a mixture G of a finite number, say g, of groups $G_1, \ldots, G_g$, in some unknown proportions $\pi_1, \ldots, \pi_g$. A prime is used here to denote vector transpose. The mixing proportions π_i lie between zero and one and sum to unity. The jth feature vector $\boldsymbol{y}_j$ is taken to have the density $f_i(\boldsymbol{y}_j)$ in group G_i $(i = 1, \ldots, g; j = 1, \ldots, n)$. Then, unconditionally with respect to its group of origin, the jth feature vector $\boldsymbol{Y}_j$ has the mixture density

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j) \quad (j = 1, \ldots, n). \tag{1}$$

In this mixture framework, the posterior probability that the jth entity with feature vector $\boldsymbol{y}_j$ belongs to G_i is given by

$$\tau_i(\boldsymbol{y}_j) = \pi_i f_i(\boldsymbol{y}_j) / f(\boldsymbol{y}_j) \quad (i = 1, \ldots, g)$$

for $j = 1, \ldots, n$. In practice the component densities $f_i(\boldsymbol{y})$ and the mixing proportions π_i will be unknown and so will have to be estimated from the observed data

$$\boldsymbol{y}_{obs} = (\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_n)'$$

This can be undertaken parametrically by specifying the component densities up to a number of unknown parameters and using the method of maximum likelihood to estimate them.

A common assumption in practice is to take the component densities to be multivariate normal. Obvious nonnormal structure in the data such as substantial skewness in the marginal distributions as evidenced from preliminary scatter plots may be able to be removed by appropriate transformations on the highly structured variates. If this is so, one can proceed on the basis that any nonnormal features in the data are due to some underlying group structure. Frequently, the consequent clusters in the data are essentially elliptical, so that it is reasonable to consider fitting mixtures of elliptically symmetric component densities. Within this class of component densities, the multivariate normal density is a convenient choice given its computational tractability.

3 Fitting Normal Mixture Models via Maximum Likelihood

Under the assumption of multivariate normal components, the ith group-conditional density $f_i(\boldsymbol{y};\, \boldsymbol{\theta}_i)$ is given by

$$\begin{aligned} f_i(\boldsymbol{y};\, \boldsymbol{\theta}_i) \; = \; & (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \\ & \exp\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} \\ & (\boldsymbol{y} - \boldsymbol{\mu}_i)\}, \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}_i$ denotes the parameter vector consisting of the elements of $\boldsymbol{\mu}_i$ and the $\frac{1}{2}p(p+1)$ distinct elements of $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. We now write the normal mixture density $f(\boldsymbol{y})$ as $f(\boldsymbol{y}\,;\, \boldsymbol{\Psi})$, where

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}')',$$

and $\boldsymbol{\theta}$ is the vector containing the elements of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$ known *a priori* to be distinct. For example, in fitting normal mixture models, the covariance matrices in the component densities may be taken to be equal.

Under the assumption that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independent realizations of the feature vector $\boldsymbol{Y}$, the log likelihood function for $\boldsymbol{\Psi}$ is given by

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j\,;\, \boldsymbol{\theta}_i). \tag{3}$$

With the maximum likelihood approach to the estimation of $\boldsymbol{\Psi}$, an estimate is provided by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = \boldsymbol{0}. \tag{4}$$

4 Application of EM Algorithm

It is straightforward, at least in principle, to find solutions of (2) using the EM algorithm of Dempster, Laird, and Rubin (1977). For the purpose of the application of the EM algorithm, the observed data $\boldsymbol{y}_{obs}$ are regarded as being incomplete. The complete-data are taken to be the observed feature vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, along with their group-indicator vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, which are unobservable in the framework of the mixture model being fitted. The ith element z_{ij} of $\boldsymbol{z}_j$ is taken to be one or zero, according as the jth entity does or does not come from the ith group G_i $(i = 1, \ldots, g\,;\, j = 1, \ldots, n)$. The complete-data vector $\boldsymbol{x}$ is therefore given by

$$\boldsymbol{x} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')',$$

where $\boldsymbol{x}_1 = (\boldsymbol{y}_1', \boldsymbol{z}_1')', \ldots, \boldsymbol{x}_n = (\boldsymbol{y}_n', \boldsymbol{z}_n')'$ are taken to be independent and identically distributed with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ being independent realizations from a multinomial distribution consisting of one draw on g categories with respective probabilities $\pi_1, \ldots, \pi_g$. That is,

$$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \overset{iid}{\sim} \text{Mult}_g(1, \boldsymbol{\pi}),$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)'$.

For this specification, the complete-data log likelihood is

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \log\{\pi_i f_i(\boldsymbol{y}_j\,;\, \boldsymbol{\theta}_i)\}. \quad (5)$$

It is clear that the complete-data likelihood $L_c(\boldsymbol{\Psi})$ implies the incomplete-data likelihood $L(\boldsymbol{\Psi})$.

The EM algorithm is easy to program and proceeds iteratively in two steps, E (for expectation) and M (for maximization); see McLachlan and Krishnan (1996) for a recent account of the EM algorithm in a general context. On the $(k+1)$th iteration, the E-step requires the calculation of

$$Q(\boldsymbol{\Psi}\,;\, \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\{\log L_c(\boldsymbol{\Psi}) \mid \boldsymbol{y}_{obs}\},$$

the conditional expectation of the complete-data log likelihood $\log L_c(\boldsymbol{\Psi})$, given the observed data $\boldsymbol{y}_{obs}$, using the current fit $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. Since $\log L_c(\boldsymbol{\Psi})$ is a linear function of the unobservable indicator variables z_{ij}, the E-step is effected simply by replacing z_{ij} by its conditional expectation given $\boldsymbol{y}_j$, using $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. That is, z_{ij} is replaced by

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij} \mid \boldsymbol{y}_j\} &= \text{pr}_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} \\ &= \tau_i(\boldsymbol{y}_j\,;\, \boldsymbol{\Psi}^{(k)}) \end{aligned}$$
$$(i = 1, \ldots, g\,;\, j = 1, \ldots, n) \quad (6)$$

which is the current estimate of the posterior probability that the jth entity with feature vector $\boldsymbol{y}_j$ belongs to G_i $(i = 1, \ldots, g\,;\, j = 1, \ldots, n)$.

On the M-step on the $(k+1)$th iteration, the intent is to choose the value of $\boldsymbol{\Psi}$, say $\boldsymbol{\Psi}^{(k+1)}$, that maximizes

$$\begin{aligned} Q(\boldsymbol{\Psi}\,;\, \boldsymbol{\Psi}^{(k)}) &= \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j\,;\, \boldsymbol{\Psi}^{(k)}) \\ &\qquad \log\{\pi_i f_i(\boldsymbol{x}_j\,;\, \boldsymbol{\theta}_i)\}, \quad (7) \end{aligned}$$

One nice feature of the EM algorithm is that the solution to the M-step often exists in closed form, as for the present case of normal mixtures.

It follows that on the M-step of the $(k+1)$th iteration, the current fit for the mixing proportions, the component means, and the covariance matrices is given explicitly by

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})/n,$$

$$\boldsymbol{\mu}_i^{(k+1)} = \{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})\boldsymbol{y}_j\}/M_i(\boldsymbol{\Psi}^{(k)}),$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_i^{(k+1)} &= \{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)}) \\ &\qquad (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})'\}/M_i(\boldsymbol{\Psi}^{(k)}) \quad (8) \end{aligned}$$

for $i = 1, \ldots, g$, where

$$M_i(\boldsymbol{\Psi}^{(k)}) = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})$$

$$(i = 1, \ldots, g).$$

In the case where the normal components are taken to have a common covariance matrix $\mathbf{\Sigma}$, the component-specific estimate (7) is replaced by

$$\mathbf{\Sigma}^{(k+1)} = n^{-1} \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \mathbf{\Psi}^k) (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})'.$$

Another nice feature of the EM algorithm is that the mixture likelihood $L(\mathbf{\Psi})$ can never be decreased after the EM sequence. Hence

$$L(\mathbf{\Psi}^{(k+1)}) \geq L(\mathbf{\Psi}^{(k)}),$$

which implies that $L(\mathbf{\Psi}^{(k)})$ converges to some L^* for a sequence of likelihood values bounded above. The E- and M-steps are alternated repeatedly until the likelihood (or the parameter estimates) change by an arbitrarily small amount in the case of convergence. Dempster et al. (1977) showed that if the very weak condition that $Q(\mathbf{\Psi}\,;\,\mathbf{\Phi})$ is continuous in both $\mathbf{\Psi}$ and $\mathbf{\Phi}$ holds, then L^* will be a local maximum of $L(\mathbf{\Psi})$, provided the sequence is not trapped at some saddle point; see Wu (1983).

Let $\hat{\mathbf{\Psi}}$ be the chosen solution of the likelihood equation. In the absence of any information apart from the observed data, $\mathbf{\Psi}$ is usually taken to be the root of (3) corresponding to the largest of the local maxima located. That is, in those cases where $L(\mathbf{\Psi})$ has a global maximum in the interior of the parameter space, $\hat{\mathbf{\Psi}}$ is the maximum likelihood estimate of $\mathbf{\Psi}$, assuming that the global maximum has been located. The consistency of the maximum likelihood estimator for finite mixture distributions has been established under the usual regularity conditions by Redner (1982); see McLachlan and Basford (1988, Section 2.1) for further details.

Care must be taken in the choice of root of the likelihood equation in those situations where $L(\mathbf{\Psi})$ is unbounded, as for example, with mixtures of normal component densities having unequal covariance matrices. In the latter case, each data point gives rise to a singularity on the edge of the parameter space (McLachlan and Basford (1988, Section 2.1)). In such cases attention can be directed to local maxima in the interior of the parameter space. For instance, under essentially the usual regularity conditions (Kiefer (1978) and Peters and Walker (1978)), there will exist a sequence of roots of the likelihood equation that is consistent and asymptotically efficient and normally distributed. With probability tending to one, these roots correspond to local maxima in the interior of the parameter space.

In practice, however, consideration still has to be given to the problem of very large local maxima that occur as a consequence of one fitted component having a very small (but nonzero) variance or generalized variance for multivariate data. Such a component corresponds to a cluster containing a few data points very close together. For example, in fitting a mixture of two heteroscedastic normal components to some univariate data, often the largest of the local maxima is obtained by fitting that normal mixture where one component corresponds to the two points nearest to each other (and hence with a very small but nonzero variance) and the other component corresponds to the remainder of the data. This particular fit is thus not of as much practical interest as those other fits (albeit corresponding to smaller local maxima) that attempt to represent any genuine grouping that may be present in the data. There is thus a need to impose some lower bound on the size of the component variances for univariate observations and on, say, the generalized component variances for multivariate data. Hathaway (1985, 1986) has considered a constrained approach of this nature.

5 The Algorithm NMM

In order to fit a mixture model using the EM algorithm, an initial value has to be specified for the vector $\mathbf{\Psi}$ of unknown parameters, or equivalently, for the posterior probability $\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi})$ that $\boldsymbol{y}_j$ belongs to the ith component of the mixture ($i = 1, \ldots, g$; $j = 1, \ldots, n$). In a cluster analysis context it is usually more appropriate to specify the latter. Typically with the fitting of mixture models to multivariate data, the likelihood equation will have multiple roots, and so the choice of starting values of the EM algorithm requires careful consideration. In this section we describe an algorithm called NMM that we have developed, which automatically uses a selection of starting values.

The program provides starting values for the application of the EM algorithm by considering a selection obtained from three sources (a) random starts (b) hierarchical clustering-based starts, and (c) the GLVQ clustering-based starts. With (a) the user can specify the number of random starts to be used, with some provision for how they are formed. Concerning (b), the user has the option of, in either standardized or unstandardized form, the results from seven hierarchical methods (nearest neighbour, furthest neighbour, group average, median, centroid , flexible sorting , and Wards method). Method (c) is based on a clustering procedure proposed by Pal, Bezdek, and Tsao (1993) and involves a generalization of the algorithm LVQ (Learning Vector Quantization) which can be thought of as a modification of k-means. There are several algorithm parameters that the user can optionally specify; alternatively, default values are used.

The program fits the normal mixture model for each of the initial grouping specified from the three sources (a) to (c). All these computations are automatically carried out by the program. The user only has to provide the data set, the extent of the selection of the initial groupings to be used to determine starting values, and the number of components that are to be fitted. Summary information is automatically outputed for the final fit, which is taken to be the fit corresponding to the largest of the local maxima located. However, the summary information can be recovered for any distinct fit.

5.1 Program Options

As well as the options pertaining to the automatic examination of starting values covered above the following options are available.

- Whether to fit arbitrary covariance matrices or constrain the EM algorithm to fit equal covariance matrices

- Implementing the Stochastic EM algorithm in place of the normal EM algorithm (Celeux and Diebolt (1985)).

- The modification of tolerances for the stopping criteria of the EM algorithm and the maximum number of iterations to be done.

- The use of Aitkin's accelerated likelihood stopping criterion (Böhning, Dietz, Schaub, Schlattmann, and Lindsay (1994)).

- The option to supply user defined initial starting values for the EM algorithm in the form of a partition of the data, the posterior probabilities of allocation or the mixture model parameters instead of the automatic methods described above.

- To produce output files compatible with the sister program MMRESAMP which aids in determining the number of components.

These options are controlled interactively when the program is run.

The options related to starting values are read from input files at runtime or gathered interactively . If the input files are not present then default values are used.

5.2 Auxiliary Subroutines

The program utilises auxiliary subroutines CHOL (algorithm AS 6) and SYMINV (algorithm AS 7) to invert a positive definite symmetric matrix. The random number generator RANDOM (algorithm AS182) is also used. These subroutines are all available from Statlib.

5.3 Failure indications

Any subroutine where an error is possible outputs a parameter called IFAULT. When an error has occured IFAULT is allocated a number identifying the type of error, when no error has occured IFAULT is given the default value zero. How the program handles an error depends on where the error has occured. Where possible the program will record a warning and continue.

The main area where failure occurs is with the inversion of the component covariance matrices. Problems occur when the matrices are near singular.

5.4 Restrictions

The program has a number of adjustable restrictions due to constraints on the sizes of certain arrays. If these limits are unsuitable they can be increased by modifying parameter definitions at the start of each subroutine and recompiling. At present the limitations of the program are:

MNIND Max data set size=400 points

MNATT Max dimension of data set=10

MAXNG Max number of groups=10

MSTART Max number of starts held=200

MHIER Max number of Hierarchical starts=50

Also it should be noted that for extremely large data sets the hierarchical methods require a large dissimilarity matrix which may be too large to run on some systems. This can be avoided by not using the hierarchical methods.

5.5 Precision and Accuracy

Double precision has been used in the program, to achieve greater accuracy and give better protection against problems when the determinants are very close to zero. The algorithm may however be converted to single precision by changing the implicit real statements at the beginning of all subroutines to single precision. Some compilers may need changes to the intrinsic functions that are used ie abs, sqrt etc.

The program's accuracy can be affected by adjusting the stopping criteria tolerances of the EM algorithm as described in section 5.1.

5.6 Running Time and Compatibility

The running time is dependent on the dimensionality and the number of points in the data set, the number of components specified and also the number of methods used to find starting values. The program has successfully been compiled and run on various SUN machines(SUN MP630, SUN IPX, and a SUN ELC), a DEC ALPHA and a MACINTOSH Quadra 605 with reasonable running times.

6 Example

The following sample data set was provided by Dr S. Gay and consists of 204 bivariate data points. Although in this example it would be simple enough to view the data in a two-dimensional plot and make a decision on starting values for the EM algorithm, no initial plots of the data were examined until after the results for the fit had been obtained. This was done in order to mimic the situation with high dimensional data sets, where one cannot rely entirely on plots of the data in lower dimensional subspaces to produce suitable starting values.

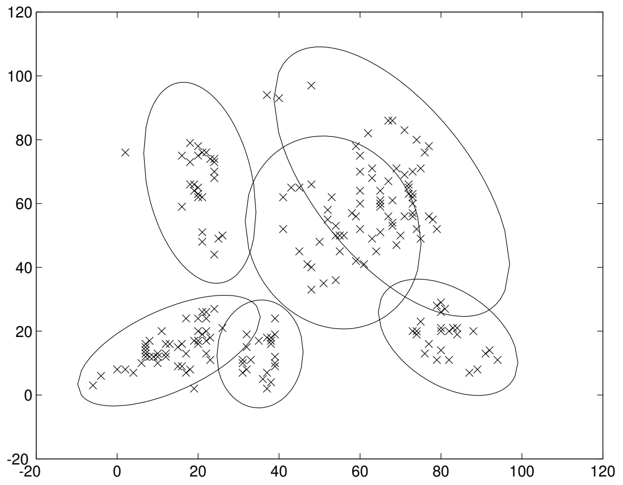

Figure 1: Initial Start Given by Ward's Method

6.1 Determining the Number of Components

A major concern in cluster analysis is deciding on the number of clusters that are present in the data, where typically as with this example data set, no *a priori* knowledge exists on the number of groups. In the context of clustering via a finite mixture model, the problem can be approached by testing for the smallest number of components in the mixture model compatible with the data. For an account of this problem, the reader is referred to McLachlan and Basford (1988, Section 1.10). More recent references may be found, for example, in Celeux and Soromenho (1996). Other approaches to this problem include those based on the minimum message length (MML) principle of Wallace and Boulton (1968), Akaike's (1973) criterion (AIC), the Bayesian information criterion (BIC) of Schwarz (1987), the approximate weight of evidence (AWE) criterion of Banfield and Raftery (1993), and the MDL principle of Rissanen (1986,1989) as applied in Rissanen and Ristad (1994). Further approaches have been proposed by Bozdogan (1992) and Windham and Cutler (1993). Also, Wallace and Dowe (1994) have recently considered the application of their SNOB program to mixture modelling using the MML principle.

The program MMRESAMP can be used to determine the most appropriate number of clusters via the resampling approach as discussed in McLachlan (1987). Basically, MMRESAMP simulates repeatedly a sample from the fitted parameters for g groups and then applies the NMM algorithm to each generated sample to fit g and $g+1$ components in turn.

The results from MMRESAMP can be seen in Table 1. The P-values lead to the decision of g=6.

Number of Groups (g)	*Log Likelihood*	$-2\log\lambda$ g *vs* $g+1$	P-*value* g *vs* $g+1$
1	-1900.81	-197.74	$< 1\%$
2	-1801.94	-97.00	$< 1\%$
3	-1753.44	-62.22	$< 1\%$
4	-1722.33	-41.80	$< 1\%$
5	-1701.43	-55.20	$< 1\%$
6	-1673.83	-16.78	13%
7	-1665.44	-8.98	33%
8	-1660.95	-16.36	18%

Table 1: Log Likelihood and P-Values

For comparison Table 2 shows the results of some other criteria for choosing the number of groups (AIC, BIC, and AWE), which lead to 11, 6, and 3 groups, respectively.

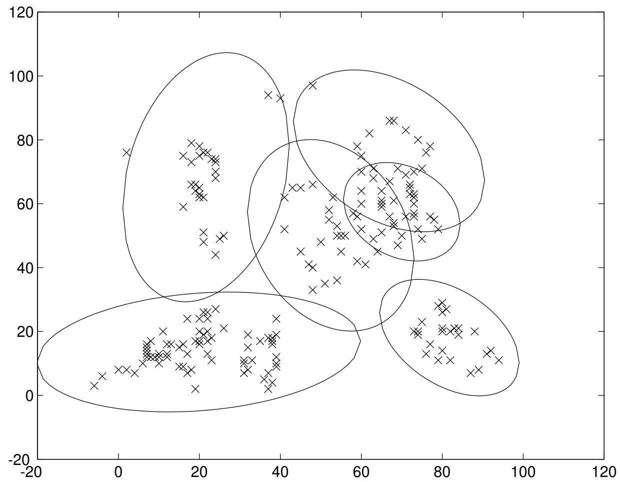

Figure 2: Initial Start Given by the GLVQ Method

Number of Groups	*AIC*	*BIC*	*AWE*
1	3811.6	3828.2	3869.8
2	3625.9	3662.4	3753.9
3	3540.9	3597.3	3738.7
4	3490.7	3567.0	3758.3
5	3460.9	3557.1	3798.3
6	3417.7	3533.8	3824.9
7	3412.9	3548.9	3890.0
8	3415.9	3571.9	3962.8
9	3411.5	3587.4	4028.3
10	3397.3	3593.0	4083.8
11	3380.4	3596.1	4136.8
12	3384.5	3620.1	4210.7

Table 2: Various Criteria

6.2 Examination of Starting Values

For our example here, starting values for the EM algorithm were obtained by using the default hierarchical methods (group average, furthest neighbour, and Ward's) each with standardized and unstandardized data), ten random starts, and the GLVQ clustering method. Table 3 reports the values of the log likelihood obtained after applying the EM algorithm for 100 iterations from each of these starts.

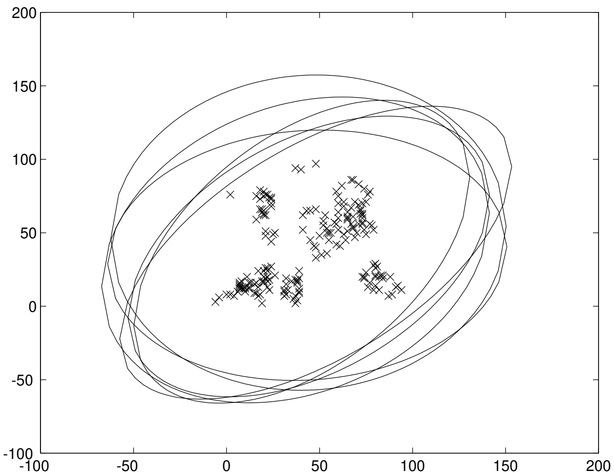

Figure 3: Initial Start Given by the Random Method

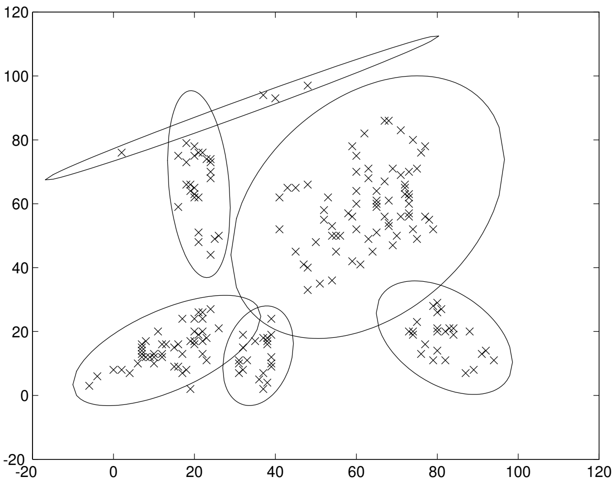

Figure 4: Fit Produced by NMM

6.3 Final Clustering

The clustering of our example data set implied by the final fit (corresponding to the largest of the local maxima located), is displayed in Figure 4. It can be seen that it has a separate cluster corresponding to four points that are almost collinear and hence for which the fitted generalized variance for that cluster is quite small relative to the generalized variances of the other clusters. As noted earlier, the solution corresponding to the largest local maximum located often has a relatively small fitted generalized variance. In this example, however, the implied clustering is of some practical consequence, as it has identified what would appear to be four outliers and put them together in a separate cluster.

The final fit discussed above was produced by a random start. However, there can be problems with the use of random starts. As points are allocated at random to each group from the whole sample this tends to give the groups similar means and variances/covariances, (especially for larger data sets) as can be seen in Figure 3. This means that the benefit from doing many random starts may be greatly reduced and an extensive number of iterations of the EM algorithm are often required to reach the final solution from these starts. With these problems in

Number of Groups g	*Log Likelihood*							
	Grp.Av. Std	*Grp.Av. Unstd*	*F.Neigh. Std*	*F.Neigh. Unstd*	*Ward's Std*	*Ward's Unstd*	*Best Random Start*	*GLVQ Start*
1	-1900.81	-1900.81	-1900.81	-1900.81	-1900.81	-1900.81	-1900.81	-1900.81
2	-1801.94	-1801.94*	-1811.63	-1811.63	-1811.63	-1811.63	-1801.94*	-1840.70
3	-1759.43	-1755.44*	-1759.43	-1759.43	-1759.43	-1759.43	-1763.39	-1759.43
4	-1725.19	-1725.19	-1725.19	-1725.19	-1722.33*	-1722.33*	-1725.19	-1725.19
5	-1717.99	-1709.42	-1705.40	-1705.40	-1709.42	-1709.42	-1701.43*	-1717.99
6	-1695.30	-1686.73	-1697.98	-1697.98	-1685.651	-1685.65	-1673.83*	-1702.68
7	-1675.29	-1679.99	-1675.29	-1675.29	-1678.89	-1665.44*	-1666.38	-1695.50
8	-1672.81	-1663.43	-1665.52	-1665.52	-1663.43	-1671.63	-1660.95*	-1686.28
9	-1656.24	-1654.96	-1657.05	-1657.05	-1654.22	-1656.16	-1652.77*	-1681.42
10	-1646.83	-1649.23	-1642.40	-1642.40	-1639.93	-1653.408	-1639.63*	-1665.19

Table 3: Log Likelihood for Various Starting Value Strategies (* Lowest Value for Given g)

mind future versions of the program will incorporate alternative methods of creating random starts which have some spatial relationship within groups.

The reason the random starts still performed quite well compared to the other starting strategies used here, is probably because 10 random starts were examined and the best chosen. Some of the 10 failed to converge on some of the fits and some gave quite bad fits.

7 Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second Inter. Symp. on Information Theory*, B.N. Petrov and F. Csaki (Eds.). Budapest: Akademiai Kiado, 267–281.

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

Bozdogan, H. (1992). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In *Information and Classification*, O.Opitz, B. Lausen, and R. Klar (Eds.). Heidelberg: Springer-Verlag, 40–54.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Instit. Statist. Math.* **46**, 373–388.

Celeux, G. and Diebolt, J. (1985). The SEM Algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Stat. Quarter.* **2**, 73–92.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *J. Classification.* To appear.

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

Hathaway, R.J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* **13**, 795–800.

Hathaway, R.J. (1986). A constrained EM algorithm for univariate mixtures. *J. Statist. Comput. Simul.* **23**, 211–230.

Kiefer, N.M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica* **46**, 427–434.

McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.* **36**, 318–324.

McLachlan (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

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

McLachlan, G.J. and Krishnan, T. (1996). *The EM Algorithm and Extensions*. New York: Wiley.

Pal, N.R., Bezdek, J.C., and Tsao, E.C.-K. (1993). Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Trans. Neural Networks* **4**, 549-557.

Peters, B.C. and Walker, H.F. (1978). An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.* **35**, 362–378.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239.

Rissanen, J. (1986). Stochastic complexity. *Ann. Statist.* **14**, 1080–1100.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Rissanen, J. and Ristad, E.S. (1994). Unsupervised classification and stochastic Complexity In *Proc. of the First US/Japan Conf. on the Frontiers of Statistical Modeling: An Informational Approach*, H. Bozdogen (Ed.). Dordrecht: Kluwer, 171–182.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

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

Wallace, C.S. and Dowe D.L. (1994). Intrinsic classification by MML - the Snob program. In 7th Australian Joint Conf. on Art. Intell., 37–44.

Windham, M.P. and Cutler, A. (1992). Information ratios for validating mixture analyses. *J. Amer. Statist. Assoc.* **87**, 1188–1192.

Wu, C.F.J.(1983). On the convergence properties of the EM algorithm. *Ann. Statist.***11**,95–103.