Computing Issues for the EM Algorithm in Mixture Models

G.J. McLachlan

Department of Mathematics
University of Queensland
Brisbane, Queensland 4072, Australia

D. Peel

Department of Mathematics
University of Queensland
Brisbane, Queensland 4072, Australia

Abstract

We consider some of the computing issues in the fitting of finite mixture models by maximum likelihood via the EM algorithm. Attention is focussed on the use of mixtures of normal components. Some of the computing issues to be addressed include the choice of an appropriate local maximizer in the case of multiple maxima, the detection of spurious local maximizers, assessment of convergence for an EM sequence, and the choice of the number of components in the mixture model. We shall describe the EMMIX algorithm for the fitting of mixture models. This algorithm automatically undertakes the fitting, including the specification of suitable initial values if not supplied by the user. The EMMIX algorithm has several options, including the provision to carry out a resampling-based test for the number of components in the mixture model and the standard errors of the fitted parameters.

1 Introduction

Finite mixtures models are being increasingly used to model the distributions of a wide variety of random phenomena; see the monographs on mixture models by Everitt and Hand (1981), Titterington, Smith, and Makov (1985), McLachlan and Basford (1988), Lindsay (1995), and Böhning (1999). The lack of homogeneity in a data set may be naturally modelled by through a mixture of distributions. Even if there is no realistic interpretation of the components of the mixture model, mixture distributions offer a very flexible modelling environment within a parametric framework. A g-component mixture model for the density function $f(\boldsymbol{y})$ of a random vector $\boldsymbol{Y}$ has the form

$$f(\boldsymbol{y}) = \sum_{i=1}^{g} \pi_1 f_i(\boldsymbol{y}), \qquad (1)$$

where $\pi_1, \ldots, \pi_g$ denote the mixing proportions which are nonnegative and sum to one, and $f_1(\boldsymbol{y}), \ldots, f_g(\boldsymbol{y})$ denote the component-density functions. Typically, the component density function $f_i(\boldsymbol{y})$ is specified up to a vector $\boldsymbol{\theta}_i$ of unknown parameters, so that then $f_i(\boldsymbol{y}) = f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$, $i = 1, \ldots, g$. In many applications, the component-density functions $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ are taken to belong to the same parametric family, for example, the normal. It can be seen from (1) that mixture models occupy an interesting niche between parametric and nonparametric approaches to statistical estimation. As explained by Jordan and Xu (1995), mixture model-based approaches are parametric in that parametric forms $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ are specified for the component density functions, but that they can also be regarded as nonparametric by allowing the number of components g to grow. Hence mixture models have much of the flexibility of nonparametric approaches, while retaining some of the advantages of parametric approaches, such as keeping the dimension of the parameter space down to a reasonable size. Mixture models therefore provide a convenient method of density estimation that lies somewhere between parametric models and kernel density estimators; see, for example, Ćwik and Koronacki (1997) and Solka et al. (1998) for some recent applications in this context.

One way of conceptualizing the mixture model (1) is to view the random vector $\boldsymbol{Y}$ as arising from the ith component of the mixture with prior probability π_1, $(i = 1, \ldots, g)$, and where the density function of $\boldsymbol{Y}$ given membership of the ith component is $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ $(i = 1, \ldots, g)$. Hence mixture models have direct applications in those situations where the random vector $\boldsymbol{Y}$ of interest can be or is to be identified as coming from one of g groups. For example, in cluster analysis applications of mixture models, the observed data are put in to g clusters by assigning each observation to its component of origin in the above conceptualization of the g-component mixture model. Here the components of the mixture model correspond to the clusters to be imposed on the data.

We let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ denote an observed p-dimensional sample of size n. The mixture model (1) can be fitted to

these data by maximum likelihood via the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1997); see also McLachlan and Krishnan (1997). The log likelihood function is given by

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i), \qquad (2)$$

where $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}^T)^T$ and $\boldsymbol{\theta}$ contains the elements of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$ known *a priori* to be distinct. With the maximum likelihood approach to the estimation of $\boldsymbol{\Psi}$, an estimate is provided by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}. \qquad (3)$$

We shall confine our attention here to the fitting of mixture models in a non-Bayesian framework. Key papers on the Bayesian analysis of mixture models include Diebolt and Robert (1994), Esobar and West (1995), Richardson and Green (1997), Robert and Mengersen (1999), and Stephens (1999); see also the papers on mixtures in Gilks et al. (1996).

As the likelihood equation (3) tends to have multiple roots for mixture models, one computing issue concerns the choice of an appropriate root. If the component-covariance matrices are unrestricted, there can be problems with spurious local maximizers. Another issue concerns determining when the EM sequence of iterates has actually converged. A further issue concerns the choice of the number of components in a mixture model. These and other issues are to be considered in the context of mixture models with normal components.

2 Normal Components

For multivariate data of a continuous nature, attention has focussed on the use of multivariate normal components because of their computational convenience. In the application of the EM algorithm, the iterates on the M-step are given in closed form. Also, in cluster analysis where a mixture model-based approach is widely adopted, the clusters in the data are often essentially elliptical in shape, 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 above-mentioned computational tractability.

We note in passing that in those situations where the tails of the normal distribution are often shorter than required, McLachlan and Peel (1998) have considered the use of mixtures of (multivariate) t distributions. The t distribution provides a longer tailed alternative to the normal distribution. Hence it provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a component are given reduced weight in the calculation of its parameters. With this t mixture model-based approach, the normal distribution for each component in the mixture is embedded in a wider class of elliptically symmetric distributions with an additional parameter called the degrees of freedom ν. As ν tends to infinity, the t distribution approaches the normal distribution. Hence this parameter ν may be viewed as a robustness tuning parameter. It can be fixed in advance or it can be inferred from the data for each component thereby providing an *adaptive* robust procedure, as explained in Lange, Little and Taylor (1989), who considered the use of a single component t distribution in linear and nonlinear regression problems.

3 Application of EM Algorithm

It is straightforward to find solutions of (3) using the EM algorithm. For the purpose of the application of the EM algorithm, the observed-data vector $\boldsymbol{y}_{obs} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ is regarded as being incomplete. As mentioned above, with a g-component mixture model, each observation $\boldsymbol{y}_j$ can be conceptualized as arising from one of the g components with probability $\pi_i (i = 1, \ldots, g)$. Corresponding to this, the component-label variables z_{ij} are consequently introduced, where z_{ij} is defined to be one or zero according as $\boldsymbol{y}_j$ did or did not arise from the ith component of the mixture model $(i = 1, \ldots, g; j = 1, \ldots, n)$. On putting $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{gj})^T$, the complete-data vector $\boldsymbol{x}_c$ is therefore given by

$$\boldsymbol{x}_c = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T,$$

where $\boldsymbol{x}_1 = (\boldsymbol{y}_1^T, \boldsymbol{z}_1^T)^T, \ldots, \boldsymbol{x}_n = (\boldsymbol{y}_n^T, \boldsymbol{z}_n^T)^T$ are 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 $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)^T$. For this specification, the complete-data log likelihood is

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \log\{\pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}. \qquad (4)$$

The EM algorithm is easy to program and proceeds iteratively in two steps, E (for expectation) and M (for

maximization). On the $(k+1)$th iteration, the E-step requires the calculation of

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

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

$$\begin{aligned}\tau_i(\boldsymbol{y}_j;\,\Psi^{(k)}) &= E_{\Psi^{(k)}}\{Z_{ij} \mid \boldsymbol{y}_j\} \\ &= {\rm pr}_{\Psi^{(k)}}\{Z_{ij}=1 \mid \boldsymbol{y}_j\} \\ &= \frac{\pi_i^{(k)}\phi(\boldsymbol{y}_j;\,\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Sigma}_i^{(k)})}{\sum_{h=1}^{g}\pi_h^{(k)}\phi(\boldsymbol{y}_j;\,\boldsymbol{\mu}_h^{(k)},\boldsymbol{\Sigma}_h^{(k)})}\end{aligned}$$

for $i=1,\ldots,g, j=1,\ldots,n$), and where $\tau_i(\boldsymbol{y}_j;\,\Psi^{(k)})$ is the estimate after the kth iteration of the posterior probability that the jth entity with feature vector $\boldsymbol{y}_j$ belongs to the ith component $(i=1,\ldots,g\,;\,j=1,\ldots,n)$.

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

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

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

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_i^{(k)}(\boldsymbol{y}_j)(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{i=1}^{n}\tau_i^{(k)}(\boldsymbol{y}_j)} \tag{5}$$

for $i=1,\ldots,g$, where $\tau_i^{(k)}(\boldsymbol{y}_j) = \tau_i(\boldsymbol{y}_j;\,\Psi^{(k)})$. An initial value has to be specified for the vector Ψ of unknown parameters for use on the E-step on the first iteration of the EM algorithm. Equivalently, initial values must be specified for the posterior probabilities of component membership of the mixture, $\tau_1(\boldsymbol{y}_j;\,\Psi^{(0)}),\ldots,\tau_g(\boldsymbol{y}_j;\,\Psi^{(0)})$, for each $\boldsymbol{y}_j\,(j=1,\ldots,n)$ for use on commencing the EM algorithm on the M-step the first time through. The latter posterior probabilities can be specified as zero-one values, corresponding to an outright classification of the data with respect to the g components of the mixture. In this case, it suffices to specify the initial partition of the data. In a cluster analysis context it is usually more appropriate to do this rather than to specify an initial value for Ψ.

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

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

which implies that $L(\Psi^{(k)})$ 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.

4 Stopping Criterion

The stopping criterion usually adopted with the EM algorithm is in terms of either the size of the relative change in the parameter estimates or the log likelihood. Böhning et al. (1994) have exploited Aitken's acceleration procedure in its application to the sequence of log likelihood values to provide a useful estimate of its limiting value. It is applicable in the case where the sequence of log likelihood values $\{l^{(k)}\}$ is linearly convergent to some value l^*, where here for brevity of notation

$$l^{(k)} = \log L(\Psi^{(k)}).$$

Under this assumption,

$$l^{(k+1)} - l^* \approx c\,(l^{(k)} - l^*), \tag{6}$$

for all k and some $c\,(0<c<1)$. The equation (6) can be rearranged to give

$$l^{(k+1)} - l^{(k)} \approx (1-c)(l^* - l^{(k)}), \tag{7}$$

for all k. It can be seen from (7) that, if c is very close to one, a small increment in the log likelihood, $l^{(k+1)}-l^{(k)}$, does not necessarily mean that $l^{(k)}$ is very close to l^*.

From (7), we have that

$$l^{(k+1)} - l^{(k)} \approx c\,(l^{(k)} - l^{(k-1)}) \tag{8}$$

for all k. Following Böhning et al. (1994), Aitken's acceleration procedure can be applied to (8) to obtain the limit l^* of the sequence of log likelihood values,

$$l^* = l^{(k)} + \frac{1}{(1-c)}(l^{(k+1)} - l^{(k)}). \tag{9}$$

Since c is unknown, it has to be estimated in (9), for example, by the ratio of successive increments,

$$c^{(k)} = (l^{(k+1)} - l^{(k)})/(l^{(k)} - l^{(k-1)}).$$

This leads to the Aitken accelerated estimate of l^*,

$$l_A^{(k+1)} = l^{(k)} + \frac{1}{(1 - c^{(k)})}(l^{(k+1)} - l^{(k)}). \qquad (10)$$

In applications where the primary interest in on the sequence of log likelihood values rather than the sequence of parameter estimates, Böhning et al. (1994) suggest the EM algorithm can be stopped if

$$\mid l_A^{(k+1)} - l_A^{(k)} \mid < \text{tol},$$

where tol is the desired tolerance. An example concerns the resampling approach (McLachlan, 1987) to the problem of assessing the null distribution of the likelihood ratio test statistic for the number of components in a mixture model, as to be discussed later. The criterion (10) is applicable for any log likelihood sequence that is linearly convergent.

5 Choice of Local Maximizer

Let $\hat{\mathbf{\Psi}}$ be the chosen solution of the likelihood equation. The likelihood function $L(\mathbf{\Psi})$ tends to have multiple local maxima for normal mixture models. In the homoscedastic case of normal components with a common covariance matrix, the likelihood function $L(\mathbf{\Psi})$ has a global maximum in the interior of the parameter space. Hence in this case in the absence of any information apart from the observed data, $\hat{\mathbf{\Psi}}$ is usually taken to be the root of (3) corresponding to the largest of the local maxima located. The consistency of the global maximizer for finite mixture distributions has been established under the usual regularity conditions; see Kiefer (1978), Peters and Walker (1978), Redner (1981) and McLachlan and Basford (1988) for further details. In the heteroscedastic case of unrestricted component covariance matrices, $L(\mathbf{\Psi})$ is unbounded, as each data point gives rise to a singularity on the edge of the parameter space. But in the heteroscedastic case, attention can be directed to local maxima in the interior of the parameter space, since under essentially the usual regularity conditions 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 has to be given to the problem of relatively large local maxima that occur as a consequence of a fitted component having a very small (but nonzero) variance for univariate data or generalized variance (the determinant of the covariance matrix) for multivariate data. Such a component corresponds to a cluster containing a few data points either relatively close together or almost lying in a lower dimensional subspace in the case of multivariate data. There is thus a need to monitor the relative size of the fitted mixing proportions and of the component variances for univariate observations and of the generalized component variances for multivariate data in an attempt to identify these spurious local maximizers. Of course the possibility here that for a given starting point the EM algorithm may converge to a spurious local maximizer or may not converge at all is not a failing of this algorithm. Rather it is a consequence of the properties of the likelihood function for the normal mixture model with unrestricted component-covariance matrices in the case of ungrouped data.

There are other ways of handling spurious local maximizers and avoiding singularities. Hathaway (1985) has considered a constrained formulation of this problem in order to avoid singularities and to reduce the number of spurious local maximizers. He showed in the case of mixtures of g-univariate normal components with variances $\sigma_1^2, \ldots, \sigma_g^2$ that by constraining the ratio of these variances $(\sigma_i^2/\sigma_h^2, i \neq h = 1, \ldots, g)$, that the constrained maximum likelihood estimator of $\mathbf{\Psi}$ is consistent, assuming that the true value of $\mathbf{\Psi}$ lies in the constrained parameter space. Another way is to adopt a penalty function that penalizes small values of the component variances (or generalized variances in the multivariate case). In a Bayesian framework, this can be effectively done through the adoption of an appropriate prior distribution that downweights small component variances or generalized variances.

6 Detection of Spurious Local Maximizers

As noted in the previous section, spurious solutions typically have a small number of points in at least one cluster which has a relatively small generalized variance. Hence the ratio of the fitted generalized component variances can be a useful guide, or warning, that a spurious solution has been found. A more informative approach is to examine the actual eigenvalues of the covariance matrix in question, rather than the determinant (which is the product of the eigenvalues). The individual eigenvalues offer a much better reflection of the clusters shape, with each eigenvalue corresponding to the variance along the elliptical axis (eigenvector) of the cluster. In this way the user can discern between small compact clusters and long thin clusters.

There is also a need to monitor the distances between the fitted component means where there appear to

be spurious local maximizers. The Euclidean distances between apparent spurious and nonspurious component means could be calculated, but may be unreliable if the feature variables are measured on disparate scales. In such cases, one may want to consider the Mahalanobis distances between the apparent spurious and nonspurious component means, using as covariance matrix an estimate for the relevant nonspurious component. Even then, small inter-component mean distances need not reflect spurious clusters, as one can have a situation where two clusters have similar means but are quite different in shape due to their having disparate covariance matrices. Hence there is really a need to monitor the distances between points in an apparent spurious cluster and the points in nearby nonspurious clusters.

To illustrate the point that a relatively small component variance does not necessarily imply a spurious solution, we consider the *Galaxy* data set analysed in Roeder (1990). This set contains measurements of the velocities of 82 galaxies diverging away from our own galaxy. In Figure 1, we give the plot of the six-component solution corresponding to the largest of the local maximizers found, along with the data in histogram form. The estimated component variances are given in Table 1.

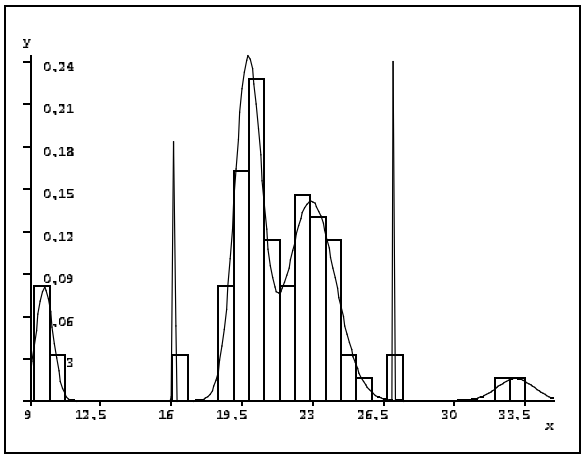

Figure 1: Plot of fitted six-component normal mixture density for Galaxy data set

From Table 1, we have a seemingly spurious solution (the two small clusters centred around 16 and 26.98), corresponding to components 2 and 5, which have relatively very small variances. However, on closer examination, it seems reasonable that the clusters in question may be legitimate, with their points not belonging to the main two clusters in the centre of the data set. This can only be confirmed if more observations become available. For this data set, Richardson and Green (1997) concluded that the number of components ranged from 5 to 7, while McLachlan and Peel (1997) provided support for $g = 6$ components.

Component	$\hat{\sigma}_i^2$
1	0.178515
2	0.001849
3	0.849564
4	1.444820
5	0.00030
6	0.454717

Table 1: Estimated component variances for the *Galaxy* data set

Another property of spurious solutions is that the EM algorithm converges to them very rarely and often only when a particular starting point is given. This reflects a very localised peak in the likelihood function. This is a tell-tale sign that a spurious solution has been found. This idea is very useful since legitimate solutions such as the one seen in the *Galaxy* data set example above will not have this property. However, for extremely large data sets, especially when there are maybe not enough clusters fitted, the repeatability aspect might not be not as useful.

7 Number of Components

In most applications of mixture models, consideration has to be given to the number of components g to be used in the mixture model. In the context of mixture models being used specifically for density estimation, the choice of the number of components g arises with consideration as to whether the mixture model has sufficient components to provide an adequate fit to the observed data. In a cluster analysis, the choice of the number of components in the mixture model relates to the question of how many clusters there are in the data. In an exploratory data analysis, as assessment of whether g is greater than one is concerned with whether any apparent separation detected in the data is a reflection of a genuine grouping or is merely due to random fluctuations in the data.

In all these contexts, it is common to approach the choice of the number of components by testing for the smallest value of g compatible with the data. This is a difficult problem, which has attracted continuing attention over the years; see, for example, Celeux and Soromenho (1996) and Biernacki, Celeux, and Govaert (1999a) for recent accounts of this problem. The likelihood ratio statistic can be used to test for the smallest number of components compatible with the data. Unfortunately, as is well known these days, regularity

conditions do not hold for the the likelihood ratio test statistic to have its usual null distribution of chi-squared with degrees of freedom equal to the difference between the number of parameters under the alternative and null hypotheses.

One approach to this problem is to adopt the resampling approach of McLachlan (1987) to assess the associated P-value. With this approach, the bootstrap is used to approximate the null distribution of the likelihood ratio statistic λ for testing $g = g_0$ versus $g = g_0 + 1$ components in the mixture model, where the value g_0 is specified by the user. Bootstrap samples are generated parametrically from the g_0-component normal mixture model with $\mathbf{\Psi}$ set equal to the fit $\hat{\mathbf{\Psi}}_{g_0}$ for $\mathbf{\Psi}$ under the null hypothesis of g_0 components. This can be carried out using an option of the EMMIX software of McLachlan et al. (1999) to be described in the next section.

In other approaches to this problem of the choice for the final value of the number of components g, several are based on the log likelihood penalized by the subtraction of some penalty term depending on the number of being parameters fitted. They include the AIC criterion of Akaike (1973), the Bayesian information criterion (BIC) of Schwarz (1987), and the approximate weight of evidence (AWE) criterion of Banfield and Raftery (1993) There are also, among others, the method of Windham and Cutler (1992) based on the rate of convergence of the EM sequence of mixture estimates and its modified version by Polymenis and Titterington (1999), the normalized entropy criterion (NEC) of Celeux and Soromenho (1996) and its modified version by Biernacki, Celeux, and Govaert (1999b), the MML principle of Wallace and Dowe (1994), and the MDL principle of Rissanen (1986,1989).

8 EMMIX Algorithm

An algorithm called EMMIX has been developed using the EM algorithm to find solutions of (1) corresponding to local maxima. In the appendix of their monograph, McLachlan and Basford (1988) gave the listing of FORTRAN programs that they had written for the maximum likelihood fitting of multivariate normal mixture models under a variety of experimental conditions. Over the years, these programs have undergone continued refinement and development, leading to an interim version known as the NMM algorithm (McLachlan and Peel, 1996). Since then, there has been much further development, culminating in the present version of the algorithm known as EMMIX (McLachlan et al., 1999).

The EMMIX algorithm automatically provides a selection of starting values for this purpose if the user does not provide any. This algorithm automatically 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) k-means clustering-based starts. Concerning (b), the user has the option of using in either standardized or unstandardized form, the results from seven hierarchical methods (nearest neighbour, farthest neighbour, group average, median, centroid, flexible sorting, and Ward's method). 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 restrictions on the component-covariance matrices (equal, unequal, or diagonal), 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 given as output for the final fit. However, it is not suggested that the clustering of a data set should be based solely on a single solution of the likelihood equation, but rather on the various solutions considered collectively. The default final fit is taken to be the one corresponding to the largest of the local maxima located. However, the summary information can be recovered for any distinct fit.

One initial criticism of the EM algorithm was that it does not automatically provide an estimate of the covariance matrix of the ML estimator, as do some other methods, such as Newton-type methods; see McLachlan and Krishnan, 1997, Chapter 4). The EMMIX algorithm has an option for the provision of standard errors for the fitted parameters in the mixture model. With this algorithm, the covariance matrix of the vector of fitted parameters $\hat{\mathbf{\Psi}}$ can be approximated either by an empirical form of the expected information matrix or by the bootstrap approach. With the latter, the bootstrap samples can be generated either parametrically from the g-component normal mixture model with $\mathbf{\Psi}$ set equal to the fit $\hat{\mathbf{\Psi}}$ or nonparametrically (that is, by sampling with replacement). Given the tendency of mixture models to have multiple local maxima at least when the sample size n is not large relative to the number of dimensions p, the bootstrap approach is favoured in this case for standard error estimation over information-based methods, which are based on a quadratic approximation to the log likelihood.

Often, in order to reduce the number of unknown parameters, the component-covariance matrices are restricted to being equal, or even diagonal as in the Au-

toClass program of Cheeseman and Stutz (1996). Less restrictive constraints can be imposed by a reparameterization of the component-covariance matrices in terms of their eigenvalue decompositions as, for example, in Banfield and Raftery (1993). In the latest version of AutoClass (http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/autoclass-c-program.html), the covariance matrices are unrestricted

Among other software for the fitting of mixture models, there are the C.A.MAN (Computer Assisted Mixture Analysis) program of Böhning, Schlattman, and Lindsay (1992) and the MIX program of Macdonald and Pitcher (1979) for univariate mixtures. More recently, there are MCLUST and EMCLUST, which are a suite of S-PLUS functions for hierarchical clustering EM, and BIC, respectively, based on parameterized normal mixture models; see Banfield and Raftery (1993) and Fraley and Raftery (1998) and the references therein. MCLUST (http://stat.washington.edu/fraley/software.shtml) and EMCLUST (http://stat.washington.edu/fraley/software.shtml) are written in FORTRAN with an interface to the S-PLUS commercial package.

Some packages for the fitting of finite mixtures have been reviewed recently by Haughton (1997). Also, Wallace and Dowe (1994) have considered the application of their SNOB (http://www.cs.monash.edu.au/ lloyd/tildeMML/Notes/SNOB.html) program to mixture modelling using the minimum message length principle of Wallace and Boulton (1968). More recently, Hunt and Jorgensen (1999) have developed the MULTIMIX program for the fitting of mixture models to data sets that contain categorical and continuous variables and which may have missing values.

9 Example

We consider now the well-known set of *Iris* data as originally collected by Anderson (1935) and first analysed by Fisher (1936). It consists of measurements of the length and width of both sepals and petals of 50 plants for each of three types of Iris species *setosa, versicolor*, and *virginica*. As pointed out by Wilson (1982), the *Iris* data were collected originally by Anderson (1935) with the view to seeing whether there was "evidence of continuing evolution in any group of plants". Her aural approach to data analysis suggested that both the *versicolor* and *virginica* species should each be split into two subspecies.

Hence we focus on the clustering of the 50 observations in the *Iris virginica* set. We considered a clustering of this data set into two clusters C_1 and C_2 by fitting a mixture of two heteroscedastic normal components. The membership of the smaller sized cluster (C_1) is reported in Table 2 for the clustering implied by each of fifteen solutions of the likelihood equation. Also listed in Table 2 for each of these local maximizers is the value of the determinant of each of the two fitted covariance matrices $|\hat{\mathbf{\Sigma}}_1|$ and $|\hat{\mathbf{\Sigma}}_2|$, and the value of the log likelihood. The clustering implied by the first solution S_1 listed in Table 2, which had been obtained previously by Wilson (1982), has nine observations in the first cluster. This solution can be found by running the EMMIX algorithm from an initial partition of the data given by either Ward's or the farthest neighbour hierarchical clustering procedures in standardized or unstandardized forms. It can also be found using random starts, but our results suggest a very large number is needed if it is to be found with a high degree of probability in a given run. It was found that these nine points in C_1 lie in an extreme portion of the scatter plot of the first two principal components using the sample correlation matrix of the data, and can be separated almost from the other points by a hyperplane. However, on the question of whether there are signs of continuing evolution in the *virginica* species, this two-group structure would not be considered significant at a conventional level. The value of $-2\log\lambda$ for the test of $g = 1$ versus $g = 2$ is 43.2. The assessed P-value, as obtained by resampling on the basis of 99 replications, is 40%. Since $g = 1$ under the null hypothesis, the bootstrap replications of $-2\log\lambda$ are actual applications of this test statistic.

In order to demonstrate the occurrence of multiple maxima, some of which may be spurious, in data sets without a strong group structure, we ran the stochastic version of the EM algorithm (Celeux, and Diebolt, 1985) from a 1000 random starting points, limiting attention to solutions at which the likelihood had a greater value than that for the first solution S_1. The stochastic version allows the EM algorithm to have a chance to escape from the current EM sequence. But evidently in this example, such escapes often led to convergence in the end to what we have concluded to be spurious local maximizers. In Table 2, we list fourteen such solutions found, labelled S_2 to S_{15}, in order of increasing value of the corresponding local maxima. They are a selection from the 51 distinct local maxima found that were greater than that for S_1 and they include the six largest found. Given the imbalance in the estimates of the component generalized variances for these fourteen solutions, they would appear to be spurious. If a lower bound were placed on each of $|\hat{\mathbf{\Sigma}}_1|$ and $|\hat{\mathbf{\Sigma}}_2|$ as discussed in Section 5, then only solution S_1 would be retained. However, it is not suggested that the clustering of a data set should be based solely on a single solution of the likelihood equation, but rather on the various solutions considered collectively. The smaller

sized clusters implied by seven of these fourteen solutions have only five members. Hence given that the data are of dimension $p = 4$, it is not surprising that the fitted covariance matrix for the first component of the mixture is nearly singular for each of them, with a generalized variance equal to only 7.6×10^{-8} or smaller. In this sense, these seven solutions would be regarded as spurious. Not withstanding that, the solutions S_2, S_4, S_6, and S_{13} provide some support to the clustering implied by S_1 in that at least four of the five members of the first cluster C_1 implied by these four solutions (actually all five members except in the case of S_6) are a subset of C_1 as implied by S_1. The first clusters C_1 implied by the solutions S_9 to S_{12} have no members in common with that of S_1. Further, it can be confirmed from scatter plots and the Mahalanobis distances between the fitted component means that solutions S_3, S_5, S_7 to S_{12}, S_{14}, and S_{15} do not provide as much separation between the means of the implied clusters. Thus these solutions would appear to be more spurious in nature rather than representing a genuine grouping.

Another way of proceeding to reduce the prevalence of solutions corresponding to artificially small values of the generalized fitted variances is to restrict the covariance matrices to be the same. Under homoscedasticity, the first cluster implied by the maximum likelihood solution (assuming it is the global maximizer), contains the union of all members of the first clusters implied by the heteroscedastic solutions S_1, S_2, S_4, S_6, and S_{13}, along with observations 9 and 36.

References

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," In *Second International Symposium on Information Theory*, B.N. Petrov and F. Csaki (Eds.), Akademiai Kiado, pp267-281.

Biernacki, C., Celeux, G., and Govaert, G. (1999a), "Assessing a mixture model for clustering with the integrated classification likelihood," *Technical Report No. 3521,* Institut National de Recherche en Informatique et en Automatique.

Biernacki, C., Celeux, G., and Govaert, G. (1999b), "An improvement of the NEC criterion for assessing the number of clusters in a mixture model," *Technical Report* Institut National de Recherche en Informatique et en Automatique.

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

Böhning, D. (1999), *Computer Assisted Analysis of Mixtures,* Chapman & Hall.

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," *Annals of the Institute of Statistical Mathematics,* Vol. 46, pp373-388.

Böhning, D., Schlattmann, P., and Lindsay, B. (1992), "Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms," *Biometrics,* Vol. 48, pp283-303.

Celeux, G., and Soromenho, G. (1996), "An entropy criterion for assessing the number of clusters in a mixture model," *Journal of Classification,* Vol. 13, pp. 195-212.

Cheeseman, P. and Stutz, J. (1996), "Bayesian classification (AutoClass): theory and results," In Advances in Knowledge Discovery and Data Mining , U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), The AAAI Press, pp. 61-83.

Ćwik, J. and Koronacki, J. (1997), "A combined adaptive-mixtures/plug-in estimator of multivariate probability densities," *Computational Statistics and Data Analysis,* Vol. 26, pp. 199-218.

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

Diebolt, J. and Robert, C.P. (1994), "Estimation of finite mixture distributions through Bayesian sampling," *Journal of the Royal Statistical Society B*, Vol. 56, pp. 363-375.

Escobar, M. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, Vol. 90, pp. 577-588.

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

Fraley, C. and Raftery, A.E. (1998), "How many clusters? Which clustering method? - Answers via model-based cluster analysis," *Computer Journal,* Vol. 41, pp. 578-588.

Gilks, W.R., Richardson, S., and Spiegelhalter D.J. (Eds.). (1996), *Markov Chain Monte Carlo in Practice,* Chapman & Hall,

Hathaway, R.J. (1985), "A constrained formulation of maximum-likelihood estimation for normal mixture distributions," *Annals of Statistics.* Vol. 13, pp. 795-800.

Hathaway, R.J. (1986), "A constrained EM algorithm for univariate normal mixtures," *Journal of Statistical Computation and Simulation,* Vol. 23, pp. 211-230.

Haughton, D. (1997), "Packages for estimating finite mixtures: a review," *The American Statistician,* Vol. 51, pp. 194-205.

Hunt, L. and Jorgensen, M.A. (1999), "Mixture model clustering using the MULTIMIX program," *Australian & New Zealand Journal of Statistics* (to appear).

Jordan, M.I. and Xu, L. (1995), "Convergence results for the EM approach to mixtures of experts architectures," *Neural Networks*, Vol. 8, pp. 1409-1431.

Kiefer, N.M. (1978), "Discrete parameter variation: Efficient estimation of a switching regression model," *Econometrika,* Vol. 46, pp. 427-434.

Lange, K., Little, R.J.A. and Taylor, J.M.G. (1989), "Robust statistical modeling using the t distribution." *Journal of the American Statistical Association,* Vol. 84, pp. 881-896.

Lindsay, B.G. (1995), *Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics,* Vol. 5, Institute of Mathematical Statistics and the American Statistical Association.

Macdonald, P. and Pitcher, (1979), "Age groups from size-frequency data: A versatile and efficient method for analyzing distribution mixtures," *Journal of the Fisheries Research Board,* Vol. 36, pp. 987-1001.

McLachlan, G.J. (1987), "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *Applied Statistics,* Vol. 36, pp. 318-324.

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

McLachlan, G.J. and Krishnan, T. (1997), *The EM Algorithm and Extensions,* Wiley.

McLachlan, G.J. and Peel, D. (1996), "An algorithm for unsupervised learning via normal mixture models," In *ISIS: Information, Statistics and Induction in Science*, D.L. Dowe, K.B. Korb, and J.J. Oliver (Eds.), pp. 354-363, World Scientific.

McLachlan, G.J. and Peel, D. (1997), Contribution to the discussion of paper by S. Richardson and P.J. Green. *Journal of the Royal Statistical Society Series B,* Vol. 59, 779-780.

McLachlan, G.J. and Peel, D. (1998), "Robust cluster analysis via mixtures of multivariate t-distributions," In *Lecture Notes in Computer Science* Vol. 1451, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.), Springer-Verlag, pp. 658-666.

McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. (1999), "The EMMIX software for the fitting of mixtures of normal and t components," *Journal of Statistical Software* (to appear).

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 Journal of Applied Mathematics,* Vol. 35, pp. 362-378.

Polymenis, A. and Titterington, D.M. (1999), "On the determination of the number of components in a mixture," *Statistics & Probability Letters.* Vol. 388, pp. 295-298.

Redner, R.A. (1981), "Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions," *Annals of Statistics.* Vol. 9, pp. 225-228.

Richardson, S., and Green, P.J. (1997), "On Bayesian analysis of mixtures with an unknown number of components"(with discussion), *Journal of the Royal Statistical Society B*, Vol. 59, pp. 731-792.

Rissanen, J. (1986), "Stochastic complexity, *Annals of Statistics,* Vol. 14, pp. 1080-1100.

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

Robert, C.P. and Mengersen K.L. (1999), "Reparameterisation issues in mixture modelling and their bearing on MCMC algorithms," *Computational Statistics and Data Analysis,* Vol. 29, pp. 325-343.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics,* Vol. 6, pp. 461-464.

Solka, J.L., Wegman, E.J., Priebe, C.E., Poston, W.L., and Rogers, G.W. (1998), "Mixture structure analysis using the Akaike information criterion and the bootstrap," *Statistics and Computing,* Vol. 8, pp. 177-188.

Stephens, M. (1999), "Dealing with multimodal posteriors and non-identifiability in mixture models," Submitted to *Journal of the Royal Statistical Society B.*

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions,* Wiley.

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

Wallace, C.S. and Dowe D.L. (1994), "Intrinsic classification by MML - the Snob program," In *7th Australian Joint Conference on Artificial Intelligence,* pp. 37-44.

Windham, M.P. and Cutler, A. (1992), "Information ratios for validating mixture analyses," *Journal of the American Statistical Association,* Vol. 87, 1188-1192.

Table 2
Results of fitting a mixture of $g = 2$
heteroscedastic normal components to data
on *Iris virginica* species

Solution No.	Cluster C_1	$\log L$	$\lvert\hat{\mathbf{\Sigma}}_1\rvert$	$\lvert\hat{\mathbf{\Sigma}}_2\rvert$
1	6,8,18,19,23, 26,30,31,32	-36.994	1.4×10^{-6}	3.7×10^{-5}
2	6,18,19,23,32	-36.987	7.6×10^{-8}	5.2×10^{-5}
3	10,13,17,21,26 30,32,40,41,42 44,46,47	-35.622	2.9×10^{-7}	7.3×10^{-5}
4	6,18,19,23,31	-35.406	6.8×10^{-9}	6.3×10^{-5}
5	5,18,21,32,35, 40,41,42,44,46	-34.427	6.2×10^{-8}	7.2×10^{-5}
6	6,18,19,32,35	-34.063	1.5×10^{-8}	5.5×10^{-5}
7	2,14,17,20,30, 32,36,43	-33.690	3.6×10^{-9}	9.7×10^{-5}
8	2,7,13,20,30 32,36,43	-32.862	3.5×10^{-10}	1.4×10^{-4}
9	1,16,41,42,45 46,49	-32.225	4.1×10^{-9}	8.8×10^{-5}
10	1,37,41,42,49	-30.374	2.9×10^{-11}	9.3×10^{-5}
11	20,35,42,46,47	-29.756	8.0×10^{-11}	8.0×10^{-5}
12	2,7,17,40,42, 43,48	-28.581	7.6×10^{-11}	1.2×10^{-4}
13	8,19,23,30,31	-27.899	8.0×10^{-11}	7.4×10^{-5}
14	8,19,23,28,39	-25.071	1.3×10^{-11}	8.0×10^{-5}
15	3,6,17,39,40	-23.536	8.9×10^{-12}	1.4×10^{-4}