On Clustering by Mixture Models

G.J. McLachlan, S.K. Ng, D. Peel

Department of Mathematics,
University of Queensland, St. Lucia, Brisbane 4072,
Australia

Abstract: Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena and to cluster data sets; see, for example, McLachlan and Peel (2000a). We consider the use of normal mixture models to cluster data sets of continuous multivariate data, concentrating on some of the associated computational issues. A robust version of this approach to clustering is obtained by modelling the data by a mixture of t distributions (Peel and McLachlan, 2000). The normal and t mixture models can be fitted by maximum likelihood via the EM algorithm, as implemented in the EMMIX software of the authors. We report some recent results of McLachlan and Ng (2000) on speeding up the fitting process by an an incremental version of the EM algorithm. The problem of clustering high-dimensional data by use of the mixture of factor analyzers model (McLachlan and Peel, 2000b) is also considered. This approach enables a normal mixture model to be fitted to data which have high dimension relative to the number of data points to be clustered.

1 Introduction

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modelling of a wide variety of random phenomena; see, for example, McLachlan and Peel (2000a). Because of their usefulness as an extremely flexible method of modelling, finite mixture models have continued to receive increasing attention over the years, both from a practical and theoretical point of view. For multivariate data of a continuous nature, attention has focussed on the use of multivariate normal components because of their wide applicability and computational convenience. They can be easily fitted iteratively by maximum likelihood (ML) via the expectation-maximization (EM) algorithm of Dempster et al. (1977); see also McLachlan and Krishnan (1997).

With a normal mixture model-based approach to clustering, it is assumed that the data to be clustered are from a mixture of an initially specified number g of multivariate normal densities in some unknown proportions $\pi_1, \ldots, \pi_g$. That is, each data point is taken to be a realization of the mixture probability density function (p.d.f.),

$$f(\boldsymbol{y};\, \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y};\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where $\phi(\boldsymbol{y};\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p-variate normal density probability function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions π_i, the elements of the component means $\boldsymbol{\mu}_i$, and the distinct elements of the component-covariance matrices $\boldsymbol{\Sigma}_i$.

Once the mixture model has been fitted, a probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data. An outright assignment of the data into g clusters is achieved by assigning each data point to the component to which it has the highest estimated posterior probability of belonging; see, for example, Bock (1996) for an account of some of the issues in cluster analysis. In this paper, we consider the use of normal and t mixture models to cluster data sets of continuous multivariate data, focussing on some of the associated computational issues.

2 Maximum likelihood estimation

2.1 Application of EM algorithm

We let $\boldsymbol{\Psi}$ be the vector of unknown parameters in the mixture model. It thus consists of the mixing proportions and the unspecified parameters in the component densities. The maximum likelihood estimate of $\boldsymbol{\Psi}$ is obtained as an appropriate root of the likelihood equation

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}, \tag{2}$$

where $L(\boldsymbol{\Psi})$ denotes the likelihood function for $\boldsymbol{\Psi}$ formed from the observed random sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Solutions of (2) corresponding to local maxima can be found by application of the EM algorithm. The latter is applied in the framework where an observation $\boldsymbol{y}_j$ is conceptualized to have arisen from one of the components and the indicator variable denoting its component of origin is taken to be missing. The E-step of the EM algorithm thus involves replacing these unobservable indicator variables by their conditional expectations given the observed data (the posterior probabilities of component membership), since the complete-data log likelihood is linear in them. For normal component densities, the estimates of their means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ can be updated in closed form on the M-step; see, for example, McLachlan and Peel (2000a, Chapter 3).

As the likelihood equation (2) tends to have multiple roots corresponding to local maxima, the EM algorithm needs to be started from a variety of initial values for the parameter vector $\boldsymbol{\Psi}$ or for a variety of initial partitions of the data into g groups. The latter can be obtained by randomly dividing the data into g groups corresponding to the g components of

the mixture model. With random starts, the effect of the central limit theorem tends to have the component parameters initially being similar at least in large samples. One way to reduce this effect is to first select a small random subsample from the data, which is then randomly assigned to the g components. The first M-step is then performed on the basis of the subsample. The subsample has to be sufficiently large to ensure that the first M-step is able to produce a nondegenerate estimate of the parameter vector $\boldsymbol{\Psi}$.

Coleman et al. (1999) have considered using a combinatorial search for a good starting point from which to apply the EM algorithm. They compared two local searches with a hierarchical agglomerative approach where the objective function to be minimized was taken to be the determinant of the pooled within-cluster covariance matrix.

2.2 Mixtures of t distributions

For many applied problems, the tails of the normal distribution are often shorter than appropriate. Also, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the normal mixture model being fitted. In this paper, we consider the fitting 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 normal component are given reduced weight in the calculation of its parameters; see McLachlan and Peel (1998), Peel and McLachlan (2000), and McLachlan and Peel (2000a, Chapter 7).

The t density with location parameter $\boldsymbol{\mu}$, positive definite matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom is given by

$$f(\boldsymbol{y};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma},\, \nu) = \frac{\Gamma(\frac{\nu+p}{2})\, |\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu}{2})\{1+\delta(\boldsymbol{y},\, \boldsymbol{\mu};\, \boldsymbol{\Sigma})/\nu\}^{\frac{1}{2}(\nu+p)}}, \tag{3}$$

where

$$\delta(\boldsymbol{y},\, \boldsymbol{\mu};\, \boldsymbol{\Sigma}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{4}$$

denotes the Mahalanobis squared distance between $\boldsymbol{y}$ and $\boldsymbol{\mu}$ (with $\boldsymbol{\Sigma}$ as the covariance matrix). If $\nu > 1$, $\boldsymbol{\mu}$ is the mean of $\boldsymbol{Y}$, and if $\nu > 2$, $\nu(\nu-2)^{-1}\boldsymbol{\Sigma}$ is its covariance matrix. As ν tends to infinity, $\boldsymbol{Y}$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The family of t distributions provides a heavy-tailed alternative to the normal family. McLachlan and Peel (2000a, Chapter 7) have provided a detailed account how the EM algorithm and a multicycle ECM variant can be used to undertake ML estimation of a mixture of t distributions.

3 Speeding up the EM algorithm

3.1 Incremental EM (IEM) algorithm

As the EM algorithm updates the posterior probabilities of component membership for each observation before the next M-step is performed, it can take some time to implement for large data sets. Hence variants of the EM algorithm have been considered for reducing the computation time. With the incremental EM (IEM) algorithm as proposed by Neal and Hinton (1998), the n available observations are divided into B ($B \leq n$) blocks and the E-step is implemented for only a block of observations at a time before the next M-step is performed. In this way, each data point $\boldsymbol{y}_j$ is visited after B partial E-steps and B (full) M-steps have been performed; that is, after one "pass" or scan of the IEM algorithm. The argument for improved rate of convergence is that the IEM algorithm exploits new information more quickly rather than waiting for a complete scan of the data before parameters are updated by an M-step. The theoretical justification for the IEM algorithm has been provided by Neal and Hinton (1998).

In implementing the M-step for normal components (or for other component densities belonging to the exponential family), it is computationally advantageous to work in terms of the current conditional expectations of the sufficient statistics. This is because the latter can be expressed partly in terms of their values on the previous iteration of the current scan and on the previous scan, so that their updating is confined effectively to a block of observations on a given E-step; see McLachlan and Peel (2000a, Chapter 12).

The choice of the number of blocks B so as to optimize the convergence time of the IEM algorithm is an interesting problem. McLachlan and Ng (2000) have investigated the tradeoff between the additional computation on one scan of the IEM algorithm and the fewer number of scans. Their results suggest using $B \approx n^{2/5}$ as a simple guide. The optimal choice will depend on the number of unknown parameters. McLachlan and Ng (2000) suggest modifying this guide to $B \approx n^{1/3}$ in the case of component-covariance matrices specified to be diagonal and to $B \approx n^{3/8}$ for component-covariance matrices restricted to be equal.

Thiesson et al. (1999) suggest a search method to choose the number of blocks. For a given number of blocks B, they propose to run the IEM algorithm for two scans (the first scan involves a full E-step) and calculate the ratio

$$r = (L_2 - L_1)/t,$$

where L_1 and L_2 are the log likelihood values after the first and the second complete scan of the data respectively, and t is the time required for the second scan which involves the partial E-step. The procedure

is repeated for various numbers of blocks B and the choice of B is the value that maximizes r.

3.2 Sparse versions of the EM and IEM algorithms

In fitting a mixture model to a data set by maximum likelihood via the EM algorithm, the current estimates of the posterior probabilities for some components of the mixture for a given data point $\boldsymbol{y}_j$ are often close to zero. Neal and Hinton (1998) proposed a sparse version of the EM algorithm for which only those component-posterior probabilities that are above a specified threshold (say, C) are updated. After running this sparse EM (SPEM) algorithm a number of iterations (say, k_1), a full EM step is then performed on which the posterior probabilities of component membership of all the observations are updated. A sparse version of the IEM algorithm (SPIEM) can be formulated by combining the sparse E-step of the SPEM algorithm and the partial E-step of the IEM algorithm.

McLachlan and Ng (2000) considered the choice of values for the threshold C and the number of iterations k_1 for the SPIEM algorithm. Their simulations suggest taking $C = 0.005$ and $k_1 = 5$. In their simulations, it was found that the use of the SPIEM algorithm reduced the time to convergence of the standard EM algorithm by a factor ranging from 18% to 71%, depending on the situation. Of course in situations where the EM algorithm is quick to converge, there may be only a little reduction in the time to convergence; indeed, in some instances, the time to convergence may actually be increased.

4 Mixtures of factor analyzers

One approach for reducing the number of unknown parameters in the forms for the component-covariance matrices $\boldsymbol{\Sigma}_i$ is to adopt the mixtures of factor analyzers model, as considered in McLachlan and Peel (2000a, 2000b). This model was originally proposed by Ghahramani and Hinton (1997) for the purposes of visualizing high dimensional data in a lower dimensional space to explore for group structure; see also Tipping and Bishop (1997) who considered the related model of mixtures of principal component analyzers for the same purpose. With the mixture of factor analyzers model, the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g), \tag{5}$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix. It assumes that the component-correlations between the observations can be explained by the conditional linear dependence of the latter on q latent or unobservable variables specific to the given component. If q is

chosen sufficiently smaller than p, the representation (5) imposes some constraints on the component-covariance matrix $\boldsymbol{\Sigma}_i$ and thus reduces the number of free parameters to be estimated.

Note that in the case of $q > 1$, there is an infinity of choices for $\boldsymbol{B}_i$, since (5) is still satisfied if $\boldsymbol{B}_i$ is replaced by $\boldsymbol{B}_i\boldsymbol{C}_i$, where $\boldsymbol{C}_i$ is any orthogonal matrix of order q. One (arbitrary) way of uniquely specifying $\boldsymbol{B}_i$ is to choose the orthogonal matrix $\boldsymbol{C}_i$ so that $\boldsymbol{B}_i^T\boldsymbol{D}_i^{-1}\boldsymbol{B}_i$ is diagonal (with its diagonal elements arranged in decreasing order); see Lawley and Maxwell (1971, Chapter 1). Assuming that the eigenvalues of $\boldsymbol{B}_i\boldsymbol{B}_i^T$ are positive and distinct, the condition that $\boldsymbol{B}_i^T\boldsymbol{D}_i^{-1}\boldsymbol{B}_i$ is diagonal as above imposes $\frac{1}{2}q(q-1)$ constraints on the parameters. Hence then the number of free parameters for each component-covariance matrix is

$$pq + p - \tfrac{1}{2}q(q-1).$$

5 Example: Clustering of microarray data

We now apply the t-mixture model to cluster some $n = 1290$ observations of $p = 26$ dimensions. These data were produced by DNA microarray experiments. $cDNA$ microarrays consist of thousands of different $cDNA$ clones spotted onto known locations on glass microscope slides. These slides/microarrays then are hybridized with differentially labelled DNA populations made from the $mRNA$s of two samples. The primary data obtained are ratios of fluorescence intensity (red/green), representing the ratios of concentrations of $mRNA$ molecules that hybridized to each of the $cDNA$s represented on the array. The data set analyzed here is a subset of the data considered in Perou et al. (2000). It contains the log ratios on 1290 gene expressions for 26 tissues on human mammary epithelial cells growing in culture and in primary human breast tumours.

In Figure 1, we display the two-cluster solution obtained by fitting a mixture of $g = 2$ t-components. The clusters are displayed in the two-dimensional space constructed using the first two principal components of the data, where we have imposed the 95% asymptotic confidence region for an observation about its component mean. The corresponding solution from fitting a two-component normal mixture model is given in Figure 2. It can be seen on comparing the two figures that the use of the t-mixture model results in a clustering that appears to be less affected by atypical observations in the data. The latter cause the estimates of the component-variances, and hence the elliptical confidence regions, to be inflated when normal components are adopted.

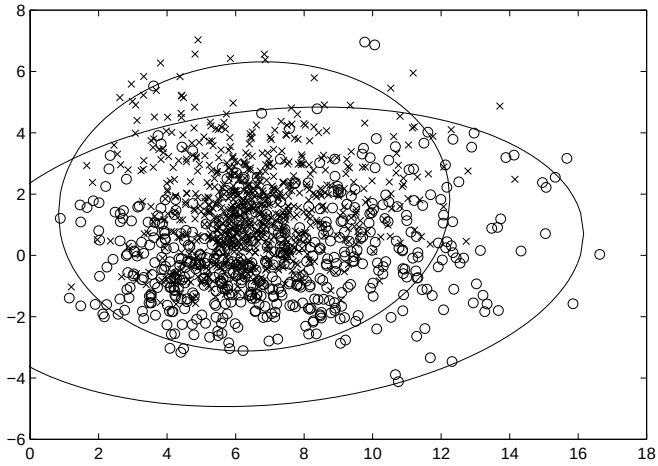

Figure 1: Plot of two-component normal mixture-based solution; (implied) cluster memberships denoted by o and x.

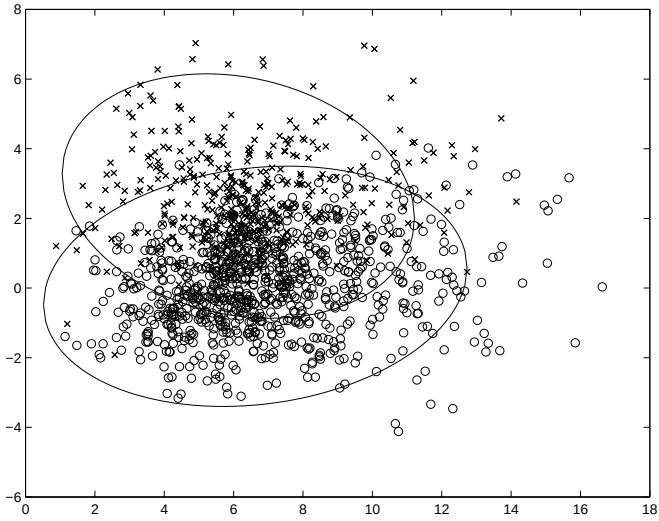

Figure 2: Plot of two-component t mixture-based solution; (implied) cluster memberships denoted by o and x.

References

BISHOP, C. M. (1998): Latent Variable Models, in Jordan (Ed.): Learning in Graphical Models, Kluwer, Dordrecht, 371-403.

BOCK, H. H. (1996): Probabilistic Models in Cluster Analysis. *Com-*

putational Statistics & Data Analysis, Vol. 23, 5–28.

COLEMAN, D., DONG, X., HARDIN, J., ROCKE, D. M., and WOODRUFF, D. L. (1999): Some Computational Issues in Cluster Analysis with No A Priori Metric. *Computational Statistics & Data Analysis, Vol. 31, 1–11.*

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, Vol. 39, 1–38.*

GHAHRAMANI, Z. and HINTON, G. E. (1997): The EM algorithm for Factor Analyzers. *Technical Report No. CRG-TR-96-1,* The University of Toronto, Toronto

McLACHLAN, G. J. and KRISHNAN, T. (1997): The EM Algorithm and Extensions. Wiley, New York

McLACHLAN, G. J. and NG, S. K. (2000): A Sparse Version of the Incremental EM Algorithm for Large Databases. *Technical Report*, Centre for Statistics, University of Queensland, Brisbane

McLACHLAN, G. J. and PEEL, D. (2000a): Finite Mixture Models. Wiley, New York

McLACHLAN, G. J. and PEEL, D. (2000b): Mixtures of Factor Analyzers, in Langley (Ed.): Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 599-606.

NEAL, R. M. and HINTON, G. E. (1998): A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants, in Jordan (Ed.): Learning in Graphical Models, Kluwer, Dordrecht, 355-368.

PEEL, D. and McLACHLAN, G. J. (2000): Robust Mixture Modelling Using the t Distribution. *Statistics & Computing, Vol. 10, 335–344.*

PEROU, C. M. et al. (2000): Distinctive Gene Expression Patterns in Human Mammary Epithelial Cells and Breast Cancers. *Proceedings of the National Academy of Sciences USA, Vol. 96, 9212–9217.*

THIESSON, B., MEEK, C., and HECKERMAN, D. (2000): Accelerating EM for Large Databases. *Technical Report No. MSR-TR-99-31* (revised version), Microsoft Research, Seattle

TIPPING, M. E. and BISHOP, C. M. (1997): Mixtures of Probabilistic Principal Component Analysers. *Technical Report No. NCRG/97/003,* Neural Computing Research Group, Aston University, Birmingham

TIPPING, M. E. and BISHOP, C. M. (1999): Mixtures of Probabilistic Principal Component Analysers. *Neural Computation, Vol. 11, 443–482.*