Clustering of High-Dimensional and Correlated Data

G.J. McLachlan, S.K. Ng and K. Wang

Abstract Finite mixture models are being commonly used in a wide range of applications in practice concerning density estimation and clustering. An attractive feature of this approach to clustering is that it provides a sound statistical framework in which to assess the important question of how many clusters there are in the data and their validity. We consider the applications of normal mixture models to high-dimensional data of a continuous nature. One way to handle the fitting of normal mixture models is to adopt mixtures of factor analyzers. However, for extremely high-dimensional data, some variable-reduction method needs to be used in conjunction with the latter model such as with the procedure called EMMIX-GENE. It was developed for the clustering of microarray data in bioinformatics, but is applicable to other types of data. We shall also consider the mixture procedure EMMIX-WIRE (based on mixtures of normal components with random effects), which is suitable for clustering high-dimensional data that may be structured (correlated and and replicated) as in longitudinal studies.

1 Introduction

Clustering procedures based on finite mixture models are being increasingly preferred over heuristic methods due to their sound mathematical basis and to the inter-

G.J. McLachlan
Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia, e-mail: gjm@maths.uq.edu.au

S.K. Ng
Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia,
e-mail: skn@maths.uq.edu.au

K. Wang
Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia,
e-mail: kwang@maths.uq.edu.au

pretability of their results. Mixture model-based procedures provide a probabilistic clustering that allows for overlapping clusters corresponding to the components of the mixture model. The uncertainties that the observations belong to the clusters are provided in terms of the fitted values for their posterior probabilities of component membership of the mixture. As each component in a finite mixture model corresponds to a cluster, it allows the important question of how many clusters there are in the data to be approached through an assessment of how many components are needed in the mixture model. These questions of model choice can be considered in terms of the likelihood function.

Scott and Symons (1971) were one of the first to adopt a model-based approach to clustering. Assuming that the data were normally distributed within a cluster, they showed that their approach is equivalent to some commonly used clustering criteria with various constraints on the cluster covariance matrices. However, from an estimation point of view, this approach yields inconsistent estimators of the parameters. This inconsistency can be avoided by working with the mixture likelihood formed under the assumption that the observed data are from a mixture of classes corresponding to the clusters to be imposed on the data, as proposed by Wolfe (1965) and Day (1969). Finite mixture models have since been increasingly used to model the distributions of a wide variety of random phenomena and to cluster data sets; see, for example, McLachlan and Peel (2000).

2 Definition of Mixture Models

We let Y denote a random vector consisting of p feature variables associated with the random phenomenon of interest. We let $y_1, \ldots, y_n$ denote an observed random sample of size n on Y. With the finite mixture model-based approach to density estimation and clustering, the density of Y is modelled as a mixture of a number (g) of component densities $f_i(y)$ 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.),

$$f(y; \Psi) = \sum_{i=1}^{g} \pi_i f_i(y), \tag{1}$$

where the mixing proportions π_i are nonnegative and sum to one. In density estimation, the number of components g can be taken sufficiently large for (1) to provide an arbitrarily accurate estimate of the underlying density function. For clustering purposes, each component in the mixture model (1) corresponds to a cluster. The posterior probability that an observation with feature vector y_j belongs to the ith component of the mixture is given by

$$\tau_i(y_j) = \pi_i f_i(y_j) / f(y_j) \tag{2}$$

for $i = 1, \ldots, g$. 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 partitioning of the observations into g (nonoverlapping) clusters $C_1, \ldots, C_g$ is effected by assigning each observation to the component to which it has the highest estimated posterior probability of belonging. Thus the ith cluster C_i contains those observations y_j with $\hat{z}_{ij} = 1$, where $\hat{z}_{ij} = 1$ if $i = h^*$, and zero otherwise, and

$$h^* = \arg\max_h \hat{\tau}_h(y_j); \tag{3}$$

$\hat{\tau}_i(y_j)$ is an estimate of $\tau_i(y_j)$. As the notation implies, $\hat{z}_{ij}$ can be viewed as an estimate of z_{ij} which, under the assumption that the observations come from a mixture of g groups $G_1, \ldots, G_g$, is defined to be one or zero according as the jth observation y_j does or does not come from G_i $(i = 1, \ldots, g;\ j = 1, \ldots, n)$.

3 Maximum Likelihood Estimation

On specifying a parametric form $f_i(y_j; \theta_i)$ for each component density, we can fit this parametric mixture model

$$f(y_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(y_j; \theta_i) \tag{4}$$

by maximum likelihood (ML). Here $\boldsymbol{\Psi} = (\omega^T, \pi_1, \ldots, \pi_{g-1})^T$ is the vector of unknown parameters, where ω consists of the elements of the θ_i known *a priori* to be distinct. In order to estimate $\boldsymbol{\Psi}$ from the observed data, it must be identifiable. This will be so if the representation (4) is unique up to a permutation of the component labels. The maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Psi}}$, is given by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = \mathbf{0}, \tag{5}$$

where $L(\boldsymbol{\Psi})$ denotes the likelihood function for $\boldsymbol{\Psi}$,

$$L(\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(y_j; \boldsymbol{\Psi}).$$

Solutions of (5) corresponding to local maximizers of $\log L(\boldsymbol{\Psi})$ can be obtained via the expectation-maximization (EM) algorithm of Dempster et al. (1977); see also McLachlan and Krishnan (1997). Let $\hat{\boldsymbol{\Psi}}$ denote the estimate of $\boldsymbol{\Psi}$ so obtained.

4 Choice of Starting Values for the EM Algorithm

McLachlan and Peel (2000) provide an in-depth account of the fitting of finite mixture models. Briefly, with mixture models the likelihood typically will have multiple maxima; that is, the likelihood equation will have multiple roots. Thus 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. Nonrandom partitions of the data can be obtained via some clustering procedure such as k-means.

The choice of root of the likelihood equation in the case of homoscedastic normal components is straightforward in the sense that the ML estimate exists as the global maximizer of the likelihood function. The situation is less straightforward in the case of heteroscedastic normal components as the likelihood function is unbounded. Usually, the intent is to choose as the ML estimate of the parameter vector $\boldsymbol{\Psi}$ the local maximizer corresponding to the largest of the local maxima located. But in practice, 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, or of the generalized component variances for multivariate data, in an attempt to identify these spurious local maximizers.

5 Clustering Via Normal Mixtures

Frequently, in practice, the 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.

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

$$f_i(\boldsymbol{y};\,\boldsymbol{\theta}_i) = \phi(\boldsymbol{y};\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{6}$$

where $\boldsymbol{\theta}_i$ consists of the elements of $\boldsymbol{\mu}_i$ and the $\frac{1}{2}p(p+1)$ distinct elements of $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. Here

$$\phi(\boldsymbol{y};\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_i|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_i)\}. \tag{7}$$

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or t-densities, is that the implied clustering is invariant under affine transformations of the data; that is, invariant under transformations of the feature vector y of the form,

$$y \to Cy + a, \tag{8}$$

where C is a nonsingular matrix. If the clustering of a procedure is invariant under (8) for only diagonal C, then it is invariant under change of measuring units but not rotations.

It can be seen from (7) that the mixture model with unrestricted component-covariance matrices in its normal component distributions is a highly parameterized one with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\Sigma_i \, (i = 1, \ldots, g)$. As an alternative to taking the component-covariance matrices to be the same or diagonal, we can adopt some model for the component-covariance matrices that is intermediate between homoscedasticity and the unrestricted model, as in the approach of Banfield and Raftery (1993). They introduced a parameterization of the component-covariance matrix Σ_i based on a variant of the standard spectral decomposition of Σ_i.

The mixture model with normal components (6) is sensitive to outliers since it adopts the multivariate normal family for the distributions of the errors. An obvious way to improve the robustness of this model for data which have longer tails than the normal or atypical observations is to consider using the multivariate t-family of elliptically symmetric distributions; see McLachlan and Peel (1998, 2000). It has an additional parameter called the degrees of freedom that controls the length of the tails of the distribution. Although the number of outliers needed for breakdown is almost the same as with the normal distribution, the outliers have to be much larger.

6 Factor Analysis Model for Dimension Reduction

As remarked earlier, the g-component normal mixture model with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\Sigma_i \, (i = 1, \ldots, g)$. As discussed above, Banfield and Raftery (1993) introduced a parameterization of the component-covariance matrix Σ_i based on a variant of the standard spectral decomposition of $\Sigma_i \, (i = 1, \ldots, g)$. However, if p is large relative to the sample size n, it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it is possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when p is large relative to n.

A common approach to reducing the number of dimensions is to perform a principal component analysis (PCA). But as is well known, projections of the feature data y_j onto the first few principal axes are not always useful in portraying the group structure. A global nonlinear approach to dimension reduction can be obtained by postulating a finite mixture of linear submodels for the distribution of the full observation vector Y_j given the (unobservable) factors. see Hinton et al. (1997), McLachlan and Peel (1998), and McLachlan et al. (2003). The mixture of factor analyzers model is given by

$$f(y_j;\Psi) = \sum_{i=1}^{g} \pi_i \phi(y_j;\mu_i,\Sigma_i), \tag{9}$$

where the ith component-covariance matrix Σ_i has the form

$$\Sigma_i = B_i B_i^T + D_i \quad (i = 1,\dots,g) \tag{10}$$

and where B_i is a $p \times q$ matrix of factor loadings and D_i is a diagonal matrix $(i = 1,\dots,g)$. The parameter vector Ψ now consists of the mixing proportions π_i and the elements of the μ_i, the B_i, and the D_i. With this approach, the number of free parameters is controlled through the dimension of the latent factor space. By working in this reduced space, it allows a model for each component-covariance matrix with complexity lying between that of the isotropic and full covariance structure models without any restrictions on the covariance matrices. The mixture of factor analyzers model can be fitted by using the alternating expectation–conditional maximization (AECM) algorithm of Meng and van Dyk (1997).

A formal test for the number of factors can be undertaken using the likelihood ratio λ, as regularity conditions hold for this test conducted at a given value for the number of components g. For the null hypothesis that $H_0 : q = q_0$ versus the alternative $H_1 : q = q_0 + 1$, the statistic $-2\log\lambda$ is asymptotically chi-squared with $d = g(p - q_0)$ degrees of freedom. However, in situations where n is not large relative to the number of unknown parameters, we prefer the use of the BIC criterion. Applied in this context, it means that twice the increase in the log likelihood $(-2\log\lambda)$ has to be greater than $d\log n$ for the null hypothesis to be rejected.

The mixture of factor analyzers model is sensitive to outliers since it uses normal errors and factors. Recently, McLachlan et al. (2007) have considered the use of mixtures of t analyzers in an attempt to make the model less sensitive to outliers.

7 Some recent extensions for high-dimensional data

The EMMIX-GENE program of McLachlan et al. (2002) has been designed for the normal mixture model-based clustering of a limited number of observations that may be of extremely high-dimensions. It was called EMIX-GENE as it was designed specifically for problems in bioinformatics that require the clustering of a relatively small number of tissue samples containing the expression levels of possibly thousands of genes. But it is applicable to clustering problems outside the field of bioinformatics involving high-dimensional data. In situations where the sample size *n* is very large relative to the dimension *p*, it might not be practical to fit mixtures of factor analyzers to data on all the variables, as it would involve a considerable amount of computation time. Thus initially some of the variables may have to be removed. Indeed, the simultaneous use of too many variables in the cluster analysis may serve only to create noise that masks the effect of a smaller number of variables. Also, the intent of the cluster analysis may not be to produce a clustering of the observations on the basis of all the available variables, but rather to discover and study different clusterings of the observations corresponding to different subsets of the variables; see, for example, Soffritti (2003) anad Galimberti and Soffritti (2007).

Therefore, the EMMIX-GENE procedure has two optional steps before the final step of clustering the observations. The first step considers the selection of a subset of relevant variables from the available set of variables by screening the variables on an individual basis to eliminate those which are of little use in clustering the observations. The usefulness of a given variable to the clustering process can be assessed formally by a test of the null hypothesis that it has a single component normal distribution over the observations. A faster but *ad hoc* way is to make this decision on the basis of the interquartile range. Even after this step has been completed, there may still remain too many variables. Thus there is a second step in EMMIX-GENE in which the retained variables are clustered (after standardization) into a number of groups on the basis of Euclidean distance so that variables with similar profiles are put into the same group. In general, care has to be taken with the scaling of variables before clustering of the observations, as the nature of the variables can be intrinsically different. Also, as noted above, the clustering of the observations via normal mixture models is invariant under changes in scale and location. The clustering of the observations can be carried out on the basis of the groups considered individually using some or all of the variables within a group or collectively. For the latter, we can replace each group by a representative (a metavariable) such as the sample mean as in the EMMIX-GENE procedure.

8 Mixtures of Normal Components with Random Effects

Up to now, we have considered the clustering of data on entities under two assumptions that are commonly adopted in practice; namely,

(a) there are no replications on any particular entity specifically identified as such;
(b) all the observations on the entities are independent of one another.

These assumptions should hold for the clustering of, say, tissue samples consisting of the expression levels of many (possibly thousands) of genes, although the tissue samples have been known to be correlated for different tissues due to flawed experimental conditions. However, condition (b) will not hold for the clustering of gene profiles, since not all the genes are independently distributed, and condition (a) will generally not hold either as the gene profiles may be measured over time or on technical replicates. While this correlated structure can be incorporated into the normal mixture model (9) by appropriate specification of the component-covariance matrices Σ_i, it is difficult to fit the model under such specifications. For example, the M-step may not exist in closed form.

Accordingly, Ng et al. (2006) have developed the procedure called EMMIX-WIRE (**EM**-based **MIX**ture analysis **W**ith **R**andom **E**ffects) to handle the clustering of correlated data that may be replicated. They adopted conditionally a mixture of linear mixed models to specify the correlation structure between the variables and to allow for correlations among the observations. It also enables covariate information to be incorporated into the clustering process.

To formulate this procedure, we consider the clustering of n gene profiles y_j $(j = 1, \ldots, n)$, where we let $y_j = (y_{1j}^T, \ldots, y_{mj}^T)^T$ contain the expression values for the jth gene profile and

$$y_{tj} = (y_{1tj}, \ldots, y_{r_t tj})^T \qquad (t = 1, \ldots, m)$$

contains the r_t replicated values in the tth biological sample $(t = 1, \ldots, m)$ on the jth gene. The dimension p of y_j is given by $p = \sum_{t=1}^{m} r_t$.

With the EMMIX-WIRE procedure, the observed p-dimensional vectors $y_1, \ldots, y_n$ are assumed to have come from a mixture of a finite number, say g, of components in some unknown proportions $\pi_1, \ldots, \pi_g$, which sum to one. Conditional on its membership of the ith component of the mixture, the profile vector y_j for the jth gene $(j = 1, \ldots, n)$, follows the model

$$y_j = X\beta_i + Ub_{ij} + Vc_i + \varepsilon_{ij}, \tag{11}$$

where the elements of β_i are fixed effects (unknown constants) modelling the conditional mean of y_j in the ith component $(i = 1, \ldots, g)$. In (11), b_{ij} (a q_b-dimensional vector) and c_i (a q_c-dimensional vector) represent the unobservable gene- and tissue-specific random effects, respectively. These random effects represent the variation due to the heterogeneity of genes and samples (corresponding to $b_i = (b_{i1}^T, \ldots, b_{in}^T)^T$ and c_i, respectively). The random effects b_i and c_i, and the measurement error vector $(\varepsilon_{i1}^T, \ldots, \varepsilon_{in}^T)^T$ are assumed to be mutually independent, where X, U, and V are known design matrices of the corresponding fixed or random effects, respectively. The presence of the random effect c_i for the expression levels of genes in the ith component induces a correlation between the profiles of genes within the same cluster.

With the LMM, the distributions of b_{ij} and c_i are taken, respectively, to be multivariate normal $N_{q_b}(0, H_i)$ and $N_{q_c}(0, \theta_{ci} I_{q_c})$, where H_i is a $q_b \times q_b$ covariance matrix and I_{q_c} is the $q_c \times q_c$ identity matrix. The measurement error vector ε_{ij} is also taken to be multivariate normal $N_p(0, A_i)$, where $A_i = \text{diag}(W\xi_i)$ is a diagonal matrix constructed from the vector $(W\xi_i)$ with $\xi_i = (\sigma_{i1}^2, \ldots, \sigma_{iq_e}^2)^T$ and W a known $p \times q_e$ zero-one design matrix.

We let $\Psi = (\psi_1^T, \ldots, \psi_g^T, \pi_1, \ldots, \pi_{g-1})^T$ be the vector of all the unknown parameters, where ψ_i is the vector containing the unknown parameters β_i, the distinct elements of H_i, θ_{ci}, and ξ_i of the ith component density $(i = 1, \ldots, g)$. The estimation of Ψ can be obtained by the ML approach via the EM algorithm, proceeding conditionally on the tissue-specific random effects c_i as formulated in Ng et al. (2006). The E- and M-steps can be implemented in closed form. In particular, an approximation to the E-step by carrying out time-consuming Monte Carlo methods is not required. A probabilistic or an outright clustering of the genes into g components can be obtained, based on the estimated posterior probabilities of component membership given the profile vectors and the estimated tissue-specific random effects $\hat{c}_i$ for $i = 1, \ldots, g$; see Ng et al. (2006).

References

1. Banfield J. and Raftery A.: Model-based gaussian and non-gaussian clustering. Biometrics. **49**, 803–821 (1993)
2. Day N.: Estimating the components of a mixture of two normal distributions. Biometrika. **56**, 463–474 (1969)
3. Dempster A., Laird N. and Rubin D.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B. **39**, 1–38 (1977)
4. Galimberti G. and Soffritti G.: Model-based methods for identifying multiple cluster structures in a data set. Computational Statistics & Data Analysis. **52**, 520–536 (2007)
5. Hinton G., Dayan P. and Revow M.: Modeling the manifolds of images of handwritten digits. IEEE Transactions on Neural Networks. **8**, 65–73 (1997)
6. McLachlan G., Bean R. and Ben-Tovim Jones L.: Extension of the mixture of factor analyzers model to incorporate the multivariate t distribution. Computational Statistics & Data Analysis. **51**, 5327–5338 (2007)
7. McLachlan G., Bean R. and Peel D.: A mixture model-based approach to the clustering of microarray expression data. Bioinformatics. **18**, 413–422 (2002)
8. McLachlan G. and Krishnan T.: The EM Algorithm and Extensions. Wiley, New York (1997)
9. McLachlan G. and Peel D.: Robust cluster analysis via mixtures of multivariate t-distributions. Lecture Notes in Computer Science, **1451**, 658–666 (1998)
10. McLachlan G. and Peel D.: Finite Mixture Models. Wiley, New York (2000)
11. McLachlan G., Peel D. and Bean R.: Modelling high-dimensional data by mixtures of factor analyzers. Computational Statistics & Data Analysis. **41**, 379–388 (2003)
12. Meng X. and van Dyk D.: The EM algorithm—an old folk song sung to a fast new tune (with discussion). Journal of the Royal Statistical Society B. **59**, 511–567 (1997)
13. Ng S., McLachlan G., Wang K., Ben-Tovim Jones L. and Ng S.: A mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics. **22**, 1745–1752 (2006)
14. Scott A. and Symons M.: Clustering methods based on likelihood ratio criteria. Biometrics. **27**, 387–397 (1971)
15. Soffritti G.: Identifying multiple cluster structures in a data matrix. Communications in Statistics - Simulation and Computation. **32**, 1151–1177 (2003)
16. Wolfe, J.: A computer program for the computation of maximum likelihood analysis of types. Technical Report SRM 65-112, U.S. Naval Personnel Research Activity, San Diego.(1965)