Mixture Model-based Statistical Pattern Recognition of Clustered or Longitudinal Data

Shu-Kay Ng and Geoffrey J. McLachlan
University of Queensland
Department of Mathematics
Brisbane QLD 4072, Australia
skn@maths.uq.edu.au
gjm@maths.uq.edu.au

Abstract

Mixture models implemented via the expectation-maximization (EM) algorithm are being increasingly used in a wide range of problems in statistical pattern recognition. For many applied problems in medical and health research, the data collected may exhibit a hierarchical structure. The independence assumption in the maximum likelihood (ML) learning of mixture models is no longer valid. Ignoring the correlation between hierarchically structured data can lead to misleading pattern recognition. In this paper, we consider the extension of Gaussian mixtures to incorporate data hierarchies via the linear mixed-effects model (LMM). Clustered and longitudinal data hierarchy settings in medical and biological research are considered.

1. Introduction

Finite mixture models have been widely applied in the field of unsupervised statistical pattern recognition, where a pattern is considered as a single entity and is represented by a finite dimensional vector of features of the pattern [6, 12]. Important applications include a variety of disciplines such as medicine, computer vision, image analysis, and machine learning; see for example [13, 15]. A common assumption in practice is to take the component densities to be Gaussian given its computational tractability. As detailed in Chapters 2 and 3 of [15], the maximum likelihood (ML) learning of Gaussian mixtures can be implemented via the expectation-maximization (EM) algorithm of [2] under the assumption of independent data.

However, for many applied problems in the context of medical, health, and biological sciences, the data collected could exhibit a hierarchical or clustered structure. Such data hierarchies may be present naturally or may be due to the experimental design. For example, in medical research, data on patients are often collected from several participating hospitals [17]. Data collected from the same hospital are often interdependent and tend to be more alike in characteristics than data chosen at random from the population as a whole. Similarly, in biological research, gene expression ratios are obtained from different tissues (patients) or there are repeated measurements of gene expression on each tissue [19, 26]. The latter is an example of longitudinal designs, where longitudinal data are obtained by a series of repeated measurements nested within individual subjects (patients). With these applications, data collected from the same unit (subject) are correlated and the independence assumption in the ML learning of Gaussian mixtures is no longer valid. Ignoring the dependence of clustered or longitudinal data can result in overlooking the importance of certain cluster or subject effects and lead to spurious or misleading pattern recognition [3].

In this paper, we consider the extension of Gaussian mixture models to incorporate data hierarchies via the linear mixed-effects model (LMM). With the LMM, cluster or subject effects are assumed to be random (random effects) and shared among data collected from the same unit (subject) [10]. Our contribution is to create a wider applicability of mixture model-based pattern recognition for medical applications with hierarchically structured data. As an illustration for the method, we consider two common data hierarchy settings in medical and biological research. In Section 3, we illustrate the analysis of clustered data with a multi-center clinical trial setting and in Section 4, the clustering of genes with repeated measurements (longitudinal data) is considered. We also show that efficient learning of the proposed mixture of LMM can still be achieved by the ML approach via the EM algorithm.

2. Gaussian Mixtures and Linear Mixed Models

With a Gaussian mixture model, the observed p-dimensional data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ are assumed to have come from a mixture of an initially specified number g of multivariate Gaussian densities in some unknown proportions $\pi_1, \ldots, \pi_g$, which sum to one. That is, each feature vector is taken to be a realization of the mixture probability density function,

$$f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \tag{1}$$

where $\phi(\boldsymbol{x}; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ denotes the p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}_h$ and covariance matrix $\boldsymbol{\Sigma}_h$. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions $\pi_1, \ldots, \pi_{g-1}$, the elements of the component means $\boldsymbol{\mu}_h$, and the distinct elements of the component-covariance matrices $\boldsymbol{\Sigma}_h$ $(h = 1, \ldots, g)$.

The EM algorithm is a popular tool for iterative ML estimation of mixture models [15]. It has a number of desirable properties including its simplicity of implementation and reliable global convergence [14, 16]. Within the EM framework, each $\boldsymbol{y}_j$ is conceptualized to have arisen from one of the g components. We let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ denote the unobservable component-indicator vectors, where the h-th element z_{hj} of $\boldsymbol{z}_j$ is taken to be one or zero according as the j-th feature vector $\boldsymbol{y}_j$ does or does not come from the h-th component. We put $\boldsymbol{z}^T = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_N^T)$ where the superscript T denotes vector transpose. The complete data is then given by $(\boldsymbol{y}, \boldsymbol{z})$. On each iteration of the EM algorithm, there are two steps called the expectation (E) step and the maximization (M) step. The E-step involves the computation of the so-called Q-function, which is the conditional expectation of the complete-data log likelihood, given the observed data $\boldsymbol{y}$ and the current estimate for $\boldsymbol{\Psi}$. The M-step updates the estimates that maximize the Q-function with respect to $\boldsymbol{\Psi}$. With Gaussian mixtures, the update of $\boldsymbol{\Psi}$ in the M-step exists in closed form [15], Chapter 3. The E- and M-steps are alternated repeatedly until convergence. A nice property of the EM algorithm is its monotonic increasing of the log likelihood at each iteration. Starting from an arbitrary initial estimate for $\boldsymbol{\Psi}$ in the parameter space, convergence is nearly always to a local maximizer, barring very bad luck in the choice of the initial starting values [14], Section 1.7. An outright or hard clustering of the data is obtained by assigning each $\boldsymbol{y}_j$ to the component of the mixture (1) to which it has the highest posterior probability of belonging, $E(z_{hj} = 1|\boldsymbol{y})$.

With LMM, cluster or subject effects are assumed to be random and shared among data collected from the same unit (subject). Let the vector $\boldsymbol{b}$ denote the random effects that occur in the data vector $\boldsymbol{y}$. The LMM specifies the mean of $\boldsymbol{y}$ conditional on the realized $\boldsymbol{b}$ as

$$E(\boldsymbol{y} \mid \boldsymbol{b}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{b}, \tag{2}$$

where elements of $\boldsymbol{\beta}$ are fixed effects (unknown constants) modeling the mean of $\boldsymbol{y}$, and $\boldsymbol{b}$ represents the unobservable random effects which have zero mean $(E(\boldsymbol{b}) = \boldsymbol{0})$ and govern the variance-covariance structure of $\boldsymbol{y}$; see for example [10]. In (2), $\boldsymbol{X}$ and $\boldsymbol{U}$ are known design matrices of the fixed effects and random effects parts, respectively. The learning of single component LMM via the EM algorithm has been described in [14], Section 5.9, where the unobservable random effects $\boldsymbol{b}$ are treated as missing data in the framework of the EM algorithm. This approach can be extended to the present context where a Gaussian mixture of LMM is to be learned.

With the use of the EM algorithm to learn mixtures of LMM, the unobservable component indicator variables $\boldsymbol{z}$ and the random effects $\boldsymbol{b}$ are both treated as missing data in the EM framework. By assuming that the random effects are normally distributed, it follows from the normal theory that the joint distribution of the complete data $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{b})$ is also a Gaussian mixture. This facilitates the implementation of the EM algorithm for the learning of mixtures of LMM, for otherwise the complete-data log likelihood cannot be evaluated in closed form; see Section 5. In this paper, we consider both clustered and longitudinal data hierarchy settings in medical and biological research as follows.

3. Clustered data: A multi-center clinical trial

With a multicenter clinical trial data structure, it is assumed that there are M participating hospitals, and within each hospital there are n_i patients $(i = 1, \ldots, M)$ involved in the study. The total number of observations is, therefore, $N = \sum_{i=1}^{M} n_i$. The objectives are to cluster the patients into subgroups based on the observations of patient's outcome y_{ij} along with the patient's characteristics $\boldsymbol{x}_{ij}$ $(j = 1, \ldots, n_i)$ and to identify risk factors on the outcome measure. For example, this clinical trial setting can be adopted to cluster patients into subgroups with different patterns of hospital length of stay [9, 18] or hospital cost [22] and to assess diagnostic criteria of some diseases [24].

For the analysis of clustered data where patients are nested within hospitals, it is assumed that the hospital (cluster) effects are random and shared among data collected from the same hospital through the corresponding linear predictors. With reference to (2), conditional on its membership of the h-th component of the Gaussian mixture, the conditional mean of y_{ij} can be expressed as

$$\mu_{hij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_h + b_{hi} \tag{3}$$

for $i = 1, \dots, M$ and $j = 1, \dots, n_i$, where $\boldsymbol{\beta}_h$ is the vector of coefficients (fixed effects) and b_{hi} represents the unobservable random effect of the i-th hospital on the h-th component mean. With (3), the first element of $\boldsymbol{x}_{ij}$ is one to account for the bias term, and the random effects b_{hi} are taken to be i.i.d. $N(0, \theta_h)$. A positive estimated random effect b_{hi} thus indicates a larger mean for the h-th component in the i-th hospital. Under this formulation, the vector of unknown parameters $\boldsymbol{\Psi}$ now consists of $\pi_1, \dots, \pi_{g-1}$, $\boldsymbol{\beta}_h$, σ_h^2, and θ_h $(h = 1, \dots, g)$, where σ_h^2 is the h-th component-variance.

3.1. Learning via the EM algorithm

Let $\boldsymbol{b}_h^T = (b_{h1}, \dots, b_{hM})$. The complete-data log likelihood is given, apart from an additive constant, by

$$\begin{aligned} \log L_c(\boldsymbol{\Psi}) &= \sum_{i=1}^{M}\sum_{j=1}^{n_i}\sum_{h=1}^{g} z_{hij} \log \pi_h \phi_{hij} \\ &\quad - \sum_{h=1}^{g} \tfrac{1}{2}\left[M \log \theta_h + \theta_h^{-1} \boldsymbol{b}_h^T \boldsymbol{b}_h\right], \end{aligned}$$

where

$$\log \phi_{hij} = -\tfrac{1}{2}\{\log \sigma_h^2 + \sigma_h^{-2}(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_h - b_{hi})^2\}$$

and $z_{hij} = 1$ if y_{ij} belongs to the h-th component or $z_{hij} = 0$ if otherwise.

On the $(k+1)$-th iteration, the E-step computes the Q-function which involves the calculation of the following conditional expectations

$$E_{\boldsymbol{\Psi}^{(k)}}(z_{hij}|\boldsymbol{y}), \quad E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{b}_h|\boldsymbol{y}), \quad E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{b}_h^T \boldsymbol{b}_h|\boldsymbol{y}). \quad (4)$$

The conditional expectations in (4) are directly obtainable as follows:

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(k)}}(z_{hij}|\boldsymbol{y}) &= \tau_{hij}^{(k)} \\ &= \frac{\pi_h^{(k)} \phi_{hij}^{(k)}}{\sum_{l=1}^{g} \pi_l^{(k)} \phi_{lij}^{(k)}}, \end{aligned} \quad (5)$$

which is the current estimated posterior probability that y_{ij} belongs to the h-th component,

$$E_{\boldsymbol{\Psi}^{(k)}}(b_{hi}|\boldsymbol{y}) = \theta_h^{(k)} \frac{\sum_{j=1}^{n_i} \tau_{hij}^{(k)}(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_h^{(k)})}{(\sum_{j=1}^{n_i} \tau_{hij}^{(k)} \theta_h^{(k)} + \sigma_h^{2\,(k)})}, \quad (6)$$

and

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{b}_h^T \boldsymbol{b}_h|\boldsymbol{y}) &= \sum_{i=1}^{M} \frac{\sigma_h^{2\,(k)} \theta_h^{(k)}}{(\sum_{j=1}^{n_i} \tau_{hij}^{(k)} \theta_h^{(k)} + \sigma_h^{2\,(k)})} \\ &\quad + \boldsymbol{b}_h^{(k)T} \boldsymbol{b}_h^{(k)}. \end{aligned} \quad (7)$$

The M-step provides the updated estimate $\boldsymbol{\Psi}^{(k+1)}$ that maximizes the Q-function with respect to $\boldsymbol{\Psi}$. It follows that

$$\pi_h^{(k+1)} = \sum_{i=1}^{M}\sum_{j=1}^{n_i} \tau_{hij}^{(k)}/N, \quad (8)$$

$$\begin{aligned} \boldsymbol{\beta}_h^{(k+1)} &= \left(\sum_{i=1}^{M}\sum_{j=1}^{n_i} \tau_{hij}^{(k)} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T\right)^{-1} \cdot \\ &\quad \left(\sum_{i=1}^{M}\sum_{j=1}^{n_i} \tau_{hij}^{(k)} \boldsymbol{x}_{ij}(y_{ij} - b_{hi}^{(k)})\right), \end{aligned} \quad (9)$$

$$\theta_h^{(k+1)} = E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{b}_h^T \boldsymbol{b}_h|\boldsymbol{y})/M, \quad (10)$$

and

$$\sigma_h^{2\,(k+1)} = \left(\sum_{i=1}^{M}\sum_{j=1}^{n_i} \tau_{hij}^{(k)} A_{hij}^{(k)}\right) / \sum_{i=1}^{M}\sum_{j=1}^{n_i} \tau_{hij}^{(k)}, \quad (11)$$

where

$$b_{hi}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}(b_{hi}|\boldsymbol{y})$$

and

$$\begin{aligned} A_{hij}^{(k)} &= (y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_h^{(k)} - b_{hi}^{(k)})^2 + \\ &\quad \frac{\sigma_h^{2\,(k)} \theta_h^{(k)}}{(\sum_{j=1}^{n_i} \tau_{hij}^{(k)} \theta_h^{(k)} + \sigma_h^{2\,(k)})}. \end{aligned}$$

3.2. A simulation study

For illustrative purposes, we here simulate some data sets of clustered data with a multicenter clinical trial data structure. It is assumed that there are $M = 10$ hospitals and within each hospital there are $n_j = 100$ patients $(j = 1, \dots, M)$. Each $\boldsymbol{x}_{ij}$ $(i = 1, \dots, 10;\ j = 1, \dots, 100)$ is a three-dimensional vector where the first element is one. A continuous bivariate vector is generated independently from the $N(\mathbf{0}, I_2)$ distribution to form the elements of $\boldsymbol{x}_{ij}$, where I_2 denotes a two dimensional identity matrix. Realizations of $\boldsymbol{Z}$ are generated in which an outcome y_{ij} has a probability of π_h of belonging to the h-th component $(h = 1, \dots, g)$. Suppose that the h-th component is determined, an outcome y_{ij} is then generated from a Gaussian $\phi(y_{ij}, \mu_{hij}, \sigma_h^2)$, with b_{hi} generated independently from the $N(0, \theta_h)$ distribution. In the simulation experiment, we consider a two-component $(g = 2)$ Gaussian mixture and assume $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\beta}_1^T = (1.0, 0.5, 0.5)$, and $\boldsymbol{\beta}_2^T = (-1.0, -0.5, 0.5)$. Two different sets of parameter values of $(\sigma_1^2, \sigma_2^2, \theta_1, \theta_2)$ are considered in the study. We repeat 10 independent simulation experiments for each set to assess the generalization performance of the proposed method. The results are presented in Table 1. For comparison, we also include the results obtained from a Gaussian mixture model with the independence assumption. It can be seen from Table 1 that the proposed mixture of LMM shows improvement in clustering the data. In addition, it is observed that the biases in the estimators of σ_1^2 and σ_2^2 are large when the dependence of clustered data is ignored in the Gaussian mixture (independent data) model.

Table 1. Simulated results for the clustered data structure.

parameters	method	error rate
$\sigma_1^2 = \sigma_2^2 = 1.0$	mixture of LMM	19.6%
$\theta_1 = \theta_2 = 1.0$	Gaussian mixture (independent data)	26.0%
$\sigma_1^2 = \sigma_2^2 = 0.5$	mixture of LMM	14.7%
$\theta_1 = \theta_2 = 1.0$	Gaussian mixture (independent data)	21.9%

4. Clustering of Genes with Repeated Measurements

In this section, we consider the clustering of genes on the basis of the genes expression-profile vector of tissue samples. As detailed in Chapter 5 of [13], the clustering of genes can be usefully employed to form a smaller number of subgroups of genes. Each subgroup of genes is represented by a single vector (a "metagene") for the subsequent clustering of the tissue samples. Another aim of clustering the genes might be to find clusters of genes that are potentially coregulated in order to search for common motifs in upstream regions of the genes in each cluster [23] and that are powerful predictor of disease outcome [7]. In recent time, gene expression microarray experiments are being carried out with replication for capturing either biological (biological replicates) or technical (technical replicates) variability in expression levels to improve the quality of inferences made from experimental studies [19, 21]. The importance of replication has been demonstrated by Lee et al. [8].

For a gene expression microarray experiment with repeated measurements, we are given, say for each i-th gene $(i = 1, \ldots, M)$, a feature vector $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}^T, \ldots, \boldsymbol{y}_{iv}^T)^T$, where v is the number of distinct tissues (patients) and

$$\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijn_{ij}})^T \quad (j = 1, \ldots, v)$$

contains the n_{ij} replications on the i-th gene from the j-th tissue. With reference to (2), it is assumed that the random effects are shared among the repeated measurements of expression on the same gene from the same biological source. Conditional on its membership of the h-th component of the Gaussian mixture, the conditional mean of y_{ijr} is expressed as

$$\mu_{hijr} = \beta_{hj} + b_{hij} \qquad (12)$$

for $i = 1, \ldots, M$, $j = 1, \ldots, v$, and $r = 1, \ldots, n_{ij}$, where b_{hij} represents the unobservable random effect of the i-th gene from the j-th tissue on the h-th component mean and is taken to be i.i.d. $N(0, \theta_{hj})$. Under this formulation, the vector of unknown parameters $\boldsymbol{\Psi}$ now consists of $\pi_1, \ldots, \pi_{g-1}$, β_{hj}, σ_{hj}^2, and θ_{hj} $(h = 1, \ldots, g;\ j = 1, \ldots, v)$.

4.1. The E- and M-steps

Apart from an additive constant, the complete-data log likelihood is given by

$$\begin{aligned} \log L_c(\boldsymbol{\Psi}) \quad = \quad & \sum_{i=1}^{M} \sum_{j=1}^{v} \sum_{h=1}^{g} z_{hi} \{\log \pi_h \phi_{hij} \\ & -\tfrac{1}{2}[\log \theta_{hj} + \theta_{hj}^{-1} b_{hij}^2]\}, \end{aligned}$$

where $z_{hi} = 1$ if $\boldsymbol{y}_i$ belongs to the h-th component or $z_{hi} = 0$ if otherwise. Here, $\log \phi_{hij}$ is given by

$$\log \phi_{hij} = -\tfrac{1}{2}\{n_{ij} \log \sigma_{hj}^2 + S_{hij}\},$$

where

$$S_{hij} = \frac{[\boldsymbol{y}_{ij} - \mathbf{1}_{n_{ij}}(\beta_{hj} + b_{hij})]^T [\boldsymbol{y}_{ij} - \mathbf{1}_{n_{ij}}(\beta_{hj} + b_{hij})]}{\sigma_{hj}^2},$$

and where $\mathbf{1}_{n_{ij}}$ is a n_{ij}-dimensional vector of ones. On the $(k+1)$-th iteration, the E-step computes

$$\begin{aligned} \tau_{hi}^{(k)} \quad &= \quad E_{\boldsymbol{\Psi}^{(k)}}(z_{hi}|\boldsymbol{y}) \\ &= \quad \frac{\pi_h^{(k)} \prod_{j=1}^{v} \phi(\boldsymbol{y}_{ij}; \mathbf{1}_{n_{ij}} \beta_{hj}^{(k)}, \boldsymbol{V}_{hij}^{(k)})}{\sum_{l=1}^{g} \pi_l^{(k)} \prod_{j=1}^{v} \phi(\boldsymbol{y}_{ij}; \mathbf{1}_{n_{ij}} \beta_{lj}^{(k)}, \boldsymbol{V}_{lij}^{(k)})}, \end{aligned} \qquad (13)$$

where $\boldsymbol{V}_{hij}$ is an $n_{ij} \times n_{ij}$ component-covariance matrix given by

$$\boldsymbol{V}_{hij} = \sigma_{hj}^2 I_{n_{ij}} + \theta_{hj} J_{n_{ij}},$$

where $J_{n_{ij}}$ is an $n_{ij} \times n_{ij}$ matrix of ones,

$$E_{\boldsymbol{\Psi}^{(k)}}(b_{hij}|\boldsymbol{y}) = \theta_{hj}^{(k)} \sum_{r=1}^{n_{ij}} \frac{(y_{ijr} - \beta_{hj}^{(k)})}{(n_{ij}\theta_{hj}^{(k)} + {\sigma_{hj}^2}^{(k)})}, \qquad (14)$$

and

$$E_{\boldsymbol{\Psi}^{(k)}}(b_{hij}^2|\boldsymbol{y}) = \frac{{\sigma_{hj}^2}^{(k)} \theta_{hj}^{(k)}}{(n_{ij}\theta_{hj}^{(k)} + {\sigma_{hj}^2}^{(k)})} + {b_{hij}^{(k)}}^2. \qquad (15)$$

The M-step updates the estimate as follows,

$$\pi_h^{(k+1)} = \sum_{i=1}^{M} \tau_{hi}^{(k)} / M, \tag{16}$$

$$\beta_{hj}^{(k+1)} = \sum_{i=1}^{M} \sum_{r=1}^{n_{ij}} \tau_{hi}^{(k)} (y_{ijr} - b_{hij}^{(k)}) / \sum_{i=1}^{M} n_{ij} \tau_{hi}^{(k)}, \tag{17}$$

$$\theta_{hj}^{(k+1)} = \sum_{i=1}^{M} \tau_{hi}^{(k)} E_{\boldsymbol{\Psi}^{(k)}} (b_{hij}^2 | \boldsymbol{y}) / \sum_{i=1}^{M} \tau_{hi}^{(k)}, \tag{18}$$

$$\sigma_{hj}^{2\ (k+1)} = \left(\sum_{i=1}^{M} \tau_{hi}^{(k)} B_{hij}^{(k)} \right) / \sum_{i=1}^{M} n_{ij} \tau_{hi}^{(k)}, \tag{19}$$

where

$$b_{hij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}} (b_{hij} | \boldsymbol{y})$$

and

$$B_{hij}^{(k)} = \sum_{r=1}^{n_{ij}} (y_{ijr} - \beta_{hj}^{(k)} - b_{hij}^{(k)})^2 + \frac{n_{ij} \sigma_{hj}^{2\ (k)} \theta_{hj}^{(k)}}{(n_{ij} \theta_{hj}^{(k)} + \sigma_{hj}^{2\ (k)})}.$$

4.2. A real example: Yeast galactose data

The data set has been used to study an integrated genomic and proteomic analyses of a systemically perturbed metabolic network [5] and is available from the online version of [26]. With the data, there are four replicate hybridizations for each cDNA array experiment. However, there are about 8% of missing data. A k-nearest neighbour $(k = 12)$ method has been adopted to impute all the missing values [26]. In our study, we work on the data set with missing values and allow the number of replicates n_{ij} to be different for each gene on each tissue sample. There are 194 genes and 20 tissues. The average number of replicates is 3.7. Our aim here is to cluster the genes based on the expression profile vector of tissue samples. The clusters so formed are then compared to the four functional categories available in the Gene Ontology (GO) listings [1]. The adjusted Rand index [4] is adopted to assess the degree of agreement between our partition and the four functional categories. The index is defined as

$$\text{adjusted Rand index } = (n_{correct} - c^*)/(n_{total} - c^*), \tag{20}$$

where $n_{correct}$ is the number of correct pairwise classifications and n_{total} is the total number of clustered pairs. In (20), c^* is a correction factor that adjusts the index so that its expected value in the case of random partition is zero [4]. It can be seen from (20) that a larger adjusted Rand index indicates a higher level of agreement. The results are presented in Table 2. For comparison, we also cluster the genes on the basis of the mean expression for each tissue. As the repeated measurements are averaged to form the mean expression profile, the information on the variability between replicates is discarded and only the information about the mean expression level utilized. It is shown in Table 2 that this approach assumes the independence of data and produces the clustering of genes that has lower adjusted Rand index.

Table 2. Adjusted Rand index (yeast galactose data).

method	adjusted Rand index
mixture of LMM	0.759
Gaussian mixture (independent data)	0.698

5. Discussion

We have described the extension of Gaussian mixture models to incorporate data hierarchies via the LMM. The applicability of the proposed method has been demonstrated in Sections 3 and 4 for the analyses of clustered and longitudinal data in medical and biological research, respectively. By assuming that the random effects are normally distributed, the EM algorithm can be adopted to perform the ML learning of mixture of LMM. Within the EM framework, the unobservable component indicator variables and the random effects are both treated as missing data. However, the EM algorithm may converge slowly where there is too much “missing information” [16], for example, when the dimension of the random effects is relatively large. In this case, some variants of the EM algorithm may be adopted to speed up the convergence; see for example [14], Section 5.9.

The EM framework developed in Sections 3 and 4 can be readily applied to calculate the residual maximum likelihood (REML) estimate. The REML method can be regarded as a method of estimation of the variance component θ by maximizing the restricted log likelihood function, which is the log likelihood obtained from a specified set of linearly independent error contrasts [20]. A discussion on the comparison between ML and REML methods for learning LMM is given in [10]. In some cases, it is shown that the REML method provides a less biased estimator for the variance component, compared to the ML estimation approach [11].

In the context of pattern recognition, it is typical to proceed on the basis that any nonnormal features in the data are due to some underlying group structure. A convenient choice for the component-densities is a Gaussian distribution given its computational tractability. In particular, the joint distribution of the complete-data also has the component-densities of a Gaussian. This facilitates the use of the EM algorithm for learning mixtures of LMM. The generalization of LMM to the generalized linear mixed model (GLMM) is essential for the analysis of non-normal data, for example discrete data. With the GLMM, the density is not necessarily assumed to be a Gaussian distribution and the mean is not necessarily taken as a linear combination of parameters as in (3) and (12). However, in this case, the complete-data log likelihood within the EM framework cannot be evaluated in closed form and has an integral with dimension equal to the number of levels of the random effects. Several procedures have been proposed in the literature, which include the methods using analytical approximation to the likelihood [11, 25] and the Monte Carlo EM algorithm, among others; see [16]. An example of EM-based approaches for the analysis of non-normal data is given in [17], where a two-component survival mixture model is adjusted for random hospital effects based on the GLMM method and the REML estimators for the variance component.

References

[1] M. Ashburner, C. A. Ball, J. A. Blake et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.*, 25(1):25–29, 2000.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. ser. B*, 39(1):1–38, 1977.

[3] H. Goldstein. *Multilevel Statistical Models (Second Edition)*. Arnold, London, 1995.

[4] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(2-3):193–218, 1985.

[5] T. Ideker, V. Thorsson, J. A. Ranish et al. Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. *Science*, 292(5518):929–934, 2001.

[6] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(1):4–38, 2000.

[7] L. Ben-Tovim Jones, S. K. Ng, C. Ambroise, K. Monico, N. Khan, and G. J. McLachlan. Use of microarray data via model-based classification in the study and prediction of survival from lung cancer. In *Methods of Microarray Data Analysis IV*, J. S. Shoemaker and S. M. Lin (Eds.). Springer, New York, 2005, pp. 163–173.

[8] M.-L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences USA*, 97(18):9834–9838, 2000.

[9] K. M. Leung, R. M. Elashoff, K. S. Rees et al. Hospital- and patient-related characteristics determining maternity length of stay: A hierarchical linear model approach. *American Journal of Public Health*, 88(3):377–381, 1998.

[10] C. E. McCulloch and S. R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, New York, 2001, Chapter 6.

[11] C. A. McGilchrist. Estimation in generalized mixed models. *J. Roy. Stat. Soc. ser. B*, 56(1):61–69, 1994.

[12] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992, Chapter 13.

[13] G. J. McLachlan, K. A. Do, and C. Ambroise. *Analyzing Microarray Gene Expression Data*. Wiley, Hobokin, New Jersey, 2004.

[14] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[15] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[16] S. K. Ng, T. Krishnan, and G. J. McLachlan. The EM Algorithm. In *Handbook of Computational Statistics Vol. 1*, J. Gentle, W. Hardle, and Y. Mori (Eds.). Springer-Verlag, New York, 2004, pp. 137–168.

[17] S. K. Ng, G. J. McLachlan, K. K. W. Yau, and A. H. Lee. Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. *Statistics in Medicine*, 23(17):2729–2744, 2004.

[18] S. K. Ng, K. K. W. Yau, and A. H. Lee. Modelling inpatient length of stay by a hierarchical mixture regression via the EM algorithm. *Mathematical and Computer Modelling*, 37(3-4):365–375, 2003.

[19] J. P. Novak, R. Sladek, and T. J. Hudson. Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, 79(1):104–113, 2002.

[20] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[21] P. Pavlidis, Q. Li, and W. S. Noble. The effect of replication on gene expression microarray experiments. *Bioinformatics*, 19(13):1620–1627, 2003.

[22] C. Quantin, E. Sauleau, P. Bolard et al. Modeling of high-cost patient distribution within renal failure diagnosis related group. *Journal of Clinical Epidemiology*, 52(3):251–258, 1999.

[23] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19(Suppl. 1):i273–i282, 2003.

[24] T. J. Thompson, P. J. Smith, and J. P. Boyle. Finite mixture models with concomitant information: Assessing diagnostic criteria for diabetes. *Applied Statistics*, 47(3):393–404, 1998.

[25] K. K. W. Yau. Multilevel models for survival analysis with random effects. *Biometrics*, 57(1):96–102, 2001.

[26] K. Y. Yeung, M. Medvedovic, and R. E. Bumgarner. Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(5):R34, 2003.