Title	Model-Based clustering of microarray expression data via latent Gaussian mixture models
Authors(s)	McNicholas, Paul D.; Murphy, Thomas Brendan
Publication date	2010-11-01
Publication information	Bioinformatics, 26 (21): 2705-2712
Publisher	Oxford University Press
Link to online version	http://dx.doi.org/10.1093/bioinformatics/btq498
Item record/more information	http://hdl.handle.net/10197/2836
Publisher's statement	This is a post-peer-review, pre-copyedit version of an article published in Bioinformatics (2010) 26 (21): 2705-2712. The definitive publisher-authenticated version [insert complete citation information here] is available online at: http://dx.doi.org/10.1093/bioinformatics/btq498.
Publisher's version (DOI)	10.1093/bioinformatics/btq498

BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–8

Model-Based Clustering of Microarray Expression Data via Latent Gaussian Mixture Models

Paul D. McNicholas [1,*] and Thomas Brendan Murphy [2]

[1]Department of Mathematics & Statistics, University of Guelph, Guelph, ON, Canada, N1G2W1.
[2]School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Ireland.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: In recent years, work has been carried out on clustering gene expression microarray data. Some approaches are developed from an algorithmic viewpoint whereas others are developed via the application of mixture models. In this paper, a family of eight mixture models which utilizes the factor analysis covariance structure is extended to twelve models and applied to gene expression microarray data. This modelling approach builds on previous work by introducing a modified factor analysis covariance structure, leading to a family of twelve mixture models, including parsimonious models. This family of models allows for the modelling of the correlation between gene expression levels even when the number of samples is small. Parameter estimation is carried out using a variant of the EM algorithm and model selection is achieved using the Bayesian information criterion. This expanded family of Gaussian mixture models , known as the EPGMM family, is then applied to two well-known gene expression data sets.

Results: The performance of the EPGMM family of models is quantified using the adjusted Rand index. This family of models gives very good performance, relative to existing popular clustering techniques, when applied to real gene expression microarray data.

Availability: The reduced, preprocessed data that were analyzed are available at www.paulmcnicholas.info

Contact: pmcnicho@uoguelph.ca

1 INTRODUCTION

1.1 Model-Based Clustering

Cluster analysis methods are used to find subgroups in a population. Clustering is of particular interest when analyzing gene expression data because it can be used to find subgroups that are well distinguished by their expression profiles. A number of clustering techniques are commonly used including agglomerative hierarchical, divisive hierarchical, k-means, k-medoids, and model-based clustering. Model-based clustering is a technique for estimating group membership based on parametric finite mixture models. The density of a parametric finite mixture model can be written

$$f(\mathbf{x} \mid \pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G) = \sum_{g=1}^{G} \pi_g r(\mathbf{x} \mid \boldsymbol{\theta}_g),$$

where $\pi_g \in [0, 1]$, such that $\sum_{g=1}^{G} \pi_g = 1$, is the probability of membership of sub-population g, and $r(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the density of a multivariate random variable $\mathbf{X}$ with parameters $\boldsymbol{\theta}_g$. Overviews of finite mixture models are given by McLachlan and Peel (2000a) and Frühwirth-Schnatter (2006).

In the model-based clustering literature, the finite Gaussian mixture model is most commonly used (examples include Fraley and Raftery, 2002; McLachlan *et al.*, 2002; McNicholas and Murphy, 2008, 2010). The density of a finite Gaussian mixture model is given by,

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \quad (1)$$

where $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian random variable $\mathbf{X}$ with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G)$. Note that the Gaussian mixture model has been used within the bioinformatics literature for purposes other than clustering: for example, McLachlan *et al.* (2006) apply a two-component mixture model to detect differential gene expression.

Gaussian mixture models offer an advantage over other commonly used approaches because the covariance structure can potentially account for correlation between expression levels within an expression profile. Consequently, these models are more flexible than k-means or hierarchical clustering which commonly use Euclidean distance when clustering. However, due to the high dimensional nature of expression data, additional structure needs to be assumed for the covariance matrices, so that the model can be fitted in high dimensional settings. The MCLUST (Fraley and Raftery, 2002) approach to model-based clustering, which utilizes eigen-decomposed covariance matrices, can only be applied to clustering expression profiles if a diagonal covariance structure is assumed; Yeung *et al.* (2001) were able to cluster genes using MCLUST but not expression profiles. By assuming a highly parsimonious but non-diagonal covariance structure, it is possible to cluster expression profiles whilst allowing for correlation between gene expressions.

In general, a structure like that given in Equation 1 can be used to model such data. Then the parameters, and hence group memberships, can be estimated using some variant of the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). The covariance matrices $\boldsymbol{\Sigma}_g$ can be decomposed to allow the construction of more parsimonious models.

*To whom correspondence should be addressed.

1.2 Parsimonious Gaussian Mixture Models

The factor analysis model (Spearman, 1904) assumes that a p-dimensional random vector $\mathbf{X}_i$ can be modeled using a q-dimensional vector of latent factors $\mathbf{U}_i$, where $q \ll p$. The model can be written $\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor weights, the latent variables $\mathbf{U}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix. Therefore, the marginal distribution of $\mathbf{X}_i$ is $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})$.

To illustrate the implications of the covariance matrix attached to this marginal distribution, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, suppose that X_{ij} and X_{ik} are expression levels from a sample $\mathbf{X}_i$. Then, $\text{Cov}(X_{ij}, X_{ik}) = \sigma_{jk} = \sum_{s=1}^{q} \lambda_{js}\lambda_{ks}$ for $j \neq k$, and $\text{Var}(X_{ij}) = \sigma_{jj} = \sum_{s=1}^{q} \lambda_{js}^2 + \psi_{qq}$. Hence, the matrix $\boldsymbol{\Lambda}$ models the covariance between expression levels, and a combination of the $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ matrices models the variance of expression levels. The factor analysis model allows for the modelling of a high-dimensional non-diagonal covariance matrix with a low number of parameters.

Ghahramani and Hinton (1997) proposed a mixture of factor analyzers model given by the finite Gaussian mixture model in Equation 1, with $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$. McLachlan and Peel (2000b) used the more general covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. Tipping and Bishop (1999) proposed the mixtures of probabilistic principal component analyzers model, for which the component covariance matrix is $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g\mathbf{I}_p$.

McNicholas and Murphy (2008) further generalized the factor analysis covariance structure by including the possibility of imposing the constraints: $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$ and $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_p$. The result of imposing, or not, each of these three constraints is the family of eight parsimonious Gaussian mixture models (PGMMs) that are described in Table 1. Each member of this family of models has a number of covariance parameters that is linear in data-dimensionality. This is one of the reasons that this family of models is particularly well suited to the analysis of high-dimensional data. The constraints allow for assuming common structure in the component covariance matrix $\boldsymbol{\Sigma}_g$, if appropriate. By assuming common covariance structure, a more parsimonious model can be used and this can be estimated in a more stable manner.

Table 1. The covariance structure of each parsimonious Gaussian mixture model — note that the UCU, UUC, and UUU models previously existed under different names, as described in Section 1.2.

$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	$\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$	Isotropic	Covariance Structure
C	C	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi\mathbf{I}_p$
C	C	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$
C	U	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi_g\mathbf{I}_p$
C	U	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g$
U	C	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi\mathbf{I}_p$
U	C	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$
U	U	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g\mathbf{I}_p$
U	U	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$

C = constrained, U = unconstrained.

The PGMM family has another significant advantage that is particularly important in applications involving high-dimensional data. When running the alternating expectation-conditional maximization (AECM) algorithm (Meng and van Dyk, 1997) for these models, it is advantageous to make use of the Woodbury identity (Woodbury, 1950) to avoid inverting any non-diagonal $p \times p$ matrices. Given an $n \times n$ matrix $\mathbf{A}$, an $n \times k$ matrix $\mathbf{H}$, a $k \times k$ matrix $\mathbf{C}$ and a $k \times n$ matrix $\mathbf{V}$, the Woodbury identity states that

$$(\mathbf{A}+\mathbf{HCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{H}(\mathbf{C}^{-1}+\mathbf{VA}^{-1}\mathbf{H})^{-1}\mathbf{VA}^{-1}. \quad (2)$$

Setting $\mathbf{H} = \boldsymbol{\Lambda}$, $\mathbf{V} = \boldsymbol{\Lambda}'$, $\mathbf{A} = \boldsymbol{\Psi}$ and $\mathbf{C} = \mathbf{I}_q$ in Equation 2 gives

$$(\boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}')^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\mathbf{I}_q + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}. \quad (3)$$

Now, the left hand side of Equation 3 involves inversion of a $p \times p$ matrix but the right hand side leaves only diagonal and $q \times q$ matrices to be inverted. This is a major computational advantage when modelling expression data, since $q \ll p$. A related identity for the determinant of the covariance matrix is given by

$$|\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}| = |\boldsymbol{\Psi}|/|\mathbf{I}_q - \boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}\boldsymbol{\Lambda}|. \quad (4)$$

Equations 3 and 4 are used by McLachlan and Peel (2000b) for the mixtures of factor analyzers model and by McNicholas and Murphy (2008) and McNicholas *et al.* (2010) for the PGMM family.

2 METHODOLOGY

2.1 Modified Factor Analysis Covariance Structure

The factor analysis covariance structure (cf. McLachlan and Peel, 2000b) can be further parameterized by writing $\boldsymbol{\Psi}_g = \omega_g\boldsymbol{\Delta}_g$, where $\omega_g \in \mathbb{R}$ and $\boldsymbol{\Delta}_g = \text{diag}\{\delta_1, \delta_2, \ldots, \delta_p\}$ such that $|\boldsymbol{\Delta}_g| = 1$, for $g = 1, 2, \ldots, G$. The resulting covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g$ shall be known as the modified factor analysis covariance structure. Now, this covariance structure can be used within the model-based clustering framework, opening up the possibility of models that are more parsimonious than their PGMM counterparts. Specifically, constraints can be imposed on the parameters $\boldsymbol{\Lambda}_g$, ω_g and $\boldsymbol{\Delta}_g$ leading to the twelve Gaussian mixture models illustrated in Table 2. The family of models in Table 2 will be referred to as the expanded PGMM (EPGMM) family hereafter. Table 2 contains a total of four new, parsimonious, models when compared to Table 1. Notably, all twelve members of the EPGMM family have a number of covariance parameters that is linear in the dimensionality of the data. Furthermore, the identities given in equations 3 and 4 can be used for all twelve models.

2.2 Parameter Estimation for the EPGMM Family

2.2.1 Introduction Estimation of the model parameters, *via* the AECM algorithm, is analogous to that of the PGMM parameter estimation procedure described by McNicholas and Murphy (2008). The estimates for the eight pre-existing models are obtained from the PGMM estimates by writing $\boldsymbol{\Psi}_g = |\boldsymbol{\Psi}_g|^{1/g}\boldsymbol{\Psi}_g/|\boldsymbol{\Psi}_g|^{1/g}$, and then setting $\omega_g = |\boldsymbol{\Psi}_g|^{1/g}$ and $\boldsymbol{\Delta}_g = \boldsymbol{\Psi}_g/|\boldsymbol{\Psi}_g|^{1/g}$. However, the derivation of the maximum likelihood estimates of the model parameters for the new models, requires the method of Lagrange multipliers (Lagrange, 1788). Parameter estimates for the CCUU model are derived in Section 2.2.2 and derivations for the other three new models are given at the end of said section.

2.2.2 AECM Algorithm The EM algorithm is an iterative technique for finding maximum likelihood estimates when data are

Table 2. The covariance structure, number of covariance parameters and nomenclature for each member of the EPGMM family, along with the name of the equivalent member of the PGMM family where applicable.

EPGMM Nomenclature						
$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	$\boldsymbol{\Delta}_g = \boldsymbol{\Delta}$	$\omega_g = \omega$	$\boldsymbol{\Delta}_g = \mathbf{I}_p$	PGMM Equivalent	Covariance Structure	Number of Covariance Parameters
C	C	C	C	CCC	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\mathbf{I}_p$	$[pq - q(q-1)/2] + 1$
C	C	U	C	CUC	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\mathbf{I}_p$	$[pq - q(q-1)/2] + G$
U	C	C	C	UCC	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\mathbf{I}_p$	$G[pq - q(q-1)/2] + 1$
U	C	U	C	UUC	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\mathbf{I}_p$	$G[pq - q(q-1)/2] + G$
C	C	C	U	CCU	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\boldsymbol{\Delta}$	$[pq - q(q-1)/2] + p$
C	C	U	U	–	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\boldsymbol{\Delta}$	$[pq - q(q-1)/2] + [G + (p-1)]$
U	C	C	U	UCU	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}$	$G[pq - q(q-1)/2] + p$
U	C	U	U	–	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}_g$	$G[pq - q(q-1)/2] + [G + (p-1)]$
C	U	C	U	–	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\boldsymbol{\Delta}_g$	$[pq - q(q-1)/2] + [1 + G(p-1)]$
C	U	U	U	CUU	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\boldsymbol{\Delta}_g$	$[pq - q(q-1)/2] + Gp$
U	U	C	U	–	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}_g$	$G[pq - q(q-1)/2] + [1 + G(p-1)]$
U	U	U	U	UUU	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g$	$G[pq - q(q-1)/2] + Gp$

C = constrained, U = unconstrained.

incomplete, or are treated as incomplete. In the expectation step (E-step), the expected value of the complete-data log-likelihood (Q, say) is computed, where the complete-data is the missing data plus the observed data. Then in the maximization step (M-step), Q is maximized with respect to the model parameters.

In the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993), the M-step is replaced by a number of conditional maximization (CM) steps. The AECM algorithm (Meng and van Dyk, 1997) is an extension of the ECM algorithm that permits different specification of the complete-data at each stage. Extensive details on the EM algorithm and variants thereof are given by McLachlan and Krishnan (2008).

Since there are two sources of missing data for the EPGMM family, the group memberships and the latent factors, the AECM algorithm is used for parameter estimation. We shall use z_{ig} to denote the group membership of sample i, so that $z_{ig} = 1$ if sample i is in group g and $z_{ig} = 0$ otherwise. At the first stage of the algorithm, the complete-data are $(\mathbf{x}_i, z_{ig})$ and in the E-step the z_{ig} are replaced by their expected values

$$\mathbb{E}[Z_{ig} \mid \hat{\pi}_g, \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Delta}}_g, \hat{\omega}_g] = \frac{\hat{\pi}_g \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Delta}}_g, \hat{\omega}_g)}{\sum_{h=1}^{G} \hat{\pi}_h \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Lambda}}_h, \hat{\boldsymbol{\Delta}}_h, \hat{\omega}_h)},$$

to give the expected value of the complete data log-likelihood, $\mathcal{Q}_1$ say. In the interest of brevity, the expected value of Z_{ig} will be denoted $\hat{z}_{ig}$ herein. The function $\mathcal{Q}_1$ is then maximized in the CM-step to give $\hat{\boldsymbol{\mu}}_g = \sum_{i=1}^{n} \hat{z}_{ig}\mathbf{x}_i / n_g$ and $\hat{\pi}_g = n_g/n$, where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}$ and $n = \sum_{g=1}^{G} n_g$.

At the second stage, the complete-data is $(\mathbf{x}_i, z_{ig}, \mathbf{u}_{ig})$ and in the E-step the z_{ig} are replaced by $\hat{z}_{ig}$ and the sufficient statistics for the factors $\mathbf{U}_{ig}$ are replaced by

$$\mathbb{E}[\mathbf{U}_{ig} \mid \mathbf{x}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \omega_g, \boldsymbol{\Delta}_g] = \boldsymbol{\beta}_g(\mathbf{x}_i - \boldsymbol{\mu}_g),$$

$$\mathbb{E}[\mathbf{U}_{ig}\mathbf{U}_{ig}' \mid \mathbf{x}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \omega_g, \boldsymbol{\Delta}_g] = \mathbf{I}_q - \boldsymbol{\beta}_g\boldsymbol{\Lambda}_g + \boldsymbol{\beta}_g(\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\beta}_g',$$

respectively, where $\boldsymbol{\beta}_g = \boldsymbol{\Lambda}_g'(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g)^{-1}$, to give $\mathcal{Q}_2$. The CM-step at this second stage will depend on the model. Consider the CCUU model, so that $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ and $\boldsymbol{\Delta}_g = \boldsymbol{\Delta}$. In this case, the expected complete-data log-likelihood $\mathcal{Q}_2(\boldsymbol{\Lambda}, \omega_g, \boldsymbol{\Delta})$ can be written

$$C + \frac{1}{2}\sum_{g=1}^{G} n_g \Big[p\log\omega_g^{-1} + \log|\boldsymbol{\Delta}^{-1}| - \omega_g^{-1}\operatorname{tr}\{\boldsymbol{\Delta}^{-1}\mathbf{S}_g\} + 2\omega_g^{-1}\operatorname{tr}\{\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\} - \omega_g^{-1}\operatorname{tr}\{\boldsymbol{\Lambda}'\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}_g\}\Big],$$

where C is constant with respect to $\boldsymbol{\Lambda}$, ω_g and $\boldsymbol{\Delta}$, and $\boldsymbol{\Theta}_g = \mathbf{I}_q - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\beta}}_g\mathbf{S}_g\hat{\boldsymbol{\beta}}_g'$.

To maximize $\mathcal{Q}_2$ with respect to $\boldsymbol{\Lambda}$, ω_g and $\boldsymbol{\Delta}$, it is necessary to use the method of Lagrange multipliers. First, form the Lagrange function $L(\boldsymbol{\Lambda}, \omega_g, \boldsymbol{\Delta}, \kappa) = Q(\boldsymbol{\Lambda}, \omega_g, \boldsymbol{\Delta}) - \kappa(|\boldsymbol{\Delta}| - 1)$. Note that we use κ to denote the Lagrange multiplier to avoid confusion with the elements of the matrix $\boldsymbol{\Lambda}$. Differentiating L with respect to $\boldsymbol{\Lambda}$, ω_g^{-1}, $\boldsymbol{\Delta}^{-1}$ and κ, respectively, gives the following score functions.

$$S_1(\boldsymbol{\Lambda}, \omega_g, \boldsymbol{\Delta}, \kappa) = \frac{\partial L}{\partial \boldsymbol{\Lambda}} = \sum_{g=1}^{G} \frac{n_g}{\omega_g}\Big[\boldsymbol{\Delta}^{-1}\mathbf{S}_g\hat{\boldsymbol{\beta}}_g' - \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\tilde{\boldsymbol{\Theta}}_g\Big],$$

$$S_2(\boldsymbol{\Lambda}, \omega_g, \boldsymbol{\Delta}, \kappa) = \frac{\partial L}{\partial \omega_g^{-1}} = \frac{n_g}{2}\Big[p\omega_g - \operatorname{tr}\{\boldsymbol{\Delta}^{-1}\mathbf{S}_g\} + 2\operatorname{tr}\{\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\} - \operatorname{tr}\{\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}_g\boldsymbol{\Lambda}'\}\Big],$$

$$S_3(\boldsymbol{\Lambda}, \omega_g, \boldsymbol{\Delta}, \kappa) = \frac{\partial L}{\partial \boldsymbol{\Delta}^{-1}} = \frac{1}{2}\sum_{g=1}^{G} n_g\Big[\boldsymbol{\Delta} - \omega_g^{-1}\mathbf{S}_g' + 2\omega_g^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g - \omega_g^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}_g'\boldsymbol{\Lambda}'\Big] + \kappa|\boldsymbol{\Delta}|\boldsymbol{\Delta},$$

$$S_4(\boldsymbol{\Lambda}, \omega_g, \boldsymbol{\Delta}, \kappa) = \frac{\partial L}{\partial \kappa} = |\boldsymbol{\Delta}| - 1.$$

Note that S_4 is included for completeness only and solving $S_4(\boldsymbol{\Lambda}, \omega_g, \boldsymbol{\Delta}, \kappa) = 0$ just returns the constraint $|\boldsymbol{\Delta}| = 1$. Now, solving $S_1(\hat{\boldsymbol{\Lambda}}^{\text{new}}, \hat{\omega}_g, \hat{\boldsymbol{\Delta}}, \kappa) = 0$ gives

$$\hat{\boldsymbol{\Lambda}}^{\text{new}} = \left[\sum_{g=1}^{G} \frac{n_g}{\hat{\omega}_g}\mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\right]\left[\sum_{g=1}^{G} \frac{n_g}{\hat{\omega}_g}\boldsymbol{\Theta}_g\right]^{-1},$$

and solving $S_2(\hat{\boldsymbol{\Lambda}}^{\text{new}}, (\hat{\omega}_g)^{\text{new}}, \hat{\boldsymbol{\Delta}}, \kappa) = 0$ gives

$$(\hat{\omega}_g)^{\text{new}} = \frac{1}{p}\,\text{tr}\left\{\hat{\boldsymbol{\Delta}}^{-1}\left[\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\right]\right\}.$$

Solving $\text{diag}\{S_3(\hat{\boldsymbol{\Lambda}}^{\text{new}}, (\hat{\omega}_g)^{\text{new}}, \hat{\boldsymbol{\Delta}}^{\text{new}}, \kappa)\} = 0$ leads to

$$\hat{\boldsymbol{\Delta}}^{\text{new}} = \frac{1}{n+2\kappa}\,\text{diag}\left\{\sum_{g=1}^{G}\frac{n_g}{(\hat{\omega}_g)^{\text{new}}}\left[\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g'(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\right]\right\}.$$

But $\hat{\boldsymbol{\Delta}}^{\text{new}}$ is a diagonal matrix with $|\hat{\boldsymbol{\Delta}}^{\text{new}}| = 1$, therefore

$$n + 2\kappa = \left(\prod_{j=1}^{p}\xi_j\right)^{\frac{1}{p}},$$

where ξ_j is the jth element along the diagonal of the matrix

$$\sum_{g=1}^{G}\frac{n_g}{(\hat{\omega}_g)^{\text{new}}}\left[\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g'(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\right].$$

Therefore, it follows that

$$\kappa = \frac{1}{2}\left[\left(\prod_{j=1}^{p}\xi_j\right)^{\frac{1}{p}} - n\right]. \tag{5}$$

The derivations for the other three new models are similar. The estimates in the UCUU case are

$$\hat{\boldsymbol{\Lambda}}_g^{\text{new}} = \mathbf{S}_g\hat{\boldsymbol{\beta}}_g\boldsymbol{\Theta}_g^{-1},$$

$$(\hat{\omega}_g)^{\text{new}} = \frac{1}{p}\,\text{tr}\{\hat{\boldsymbol{\Delta}}^{-1}\mathbf{S}_g - \hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\},$$

$$\hat{\boldsymbol{\Delta}}^{\text{new}} = \frac{1}{n+2\kappa}\,\text{diag}\left\{\sum_{g=1}^{G}\frac{n_g}{(\hat{\omega}_g)^{\text{new}}}\left[\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\right]\right\},$$

where κ is as defined in Equation 5 but, in this case, ξ_j is the jth element along the diagonal of the matrix

$$\sum_{g=1}^{G}\frac{n_g}{(\hat{\omega}_g)^{\text{new}}}\left[\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\right].$$

In the CUCU case, the estimate for $\boldsymbol{\Lambda}$ is derived in a row-by-row fashion as

$$\hat{\boldsymbol{\lambda}}_i^{\text{new}} = \mathbf{r}_i\left(\sum_{g=1}^{G}\frac{n_g}{\hat{\delta}_{g(i)}}\boldsymbol{\Theta}_g\right)^{-1},$$

for $i = 1, \ldots, p$ where $\mathbf{r}_i$ is the ith row of the matrix $\sum_{g=1}^{G}(n_g/\hat{\delta}_{g(j)})\mathbf{S}_g\hat{\boldsymbol{\beta}}_g'$, and $\hat{\delta}_{g(i)}$ is the ith element along the diagonal of the matrix $\hat{\boldsymbol{\Delta}}_g$. The other estimates are

$$(\hat{\omega})^{\text{new}} = \frac{1}{p}\sum_{g=1}^{G}\hat{\pi}_g\,\text{tr}\left\{\hat{\boldsymbol{\Delta}}_g^{-1}\left[\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\right]\right\},$$

$$\boldsymbol{\Delta}_g^{\text{new}} = \frac{n_g}{\hat{\omega}^{\text{new}}(n_g + 2\kappa_g)}\,\text{diag}\{\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\},$$

$$\kappa_g = \frac{n_g}{2}\left[\frac{1}{(\hat{\omega}_g)^{\text{new}}}\left(\prod_{j=1}^{p}\xi_{gj}\right)^{\frac{1}{p}} - 1\right],$$

where ξ_{gj} is the jth element along the diagonal of the matrix $\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g(\hat{\boldsymbol{\Lambda}}^{\text{new}})'$.

In the CCUU case, the parameter estimates are given by

$$\hat{\boldsymbol{\Lambda}}_g^{\text{new}} = \mathbf{S}_g\hat{\boldsymbol{\beta}}_g\boldsymbol{\Theta}_g^{-1},$$

$$(\hat{\omega})^{\text{new}} = \frac{1}{p}\sum_{g=1}^{G}\hat{\pi}_g\,\text{tr}\{\hat{\boldsymbol{\Delta}}_g^{-1}(\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g)\},$$

$$\hat{\boldsymbol{\Delta}}_g^{\text{new}} = \frac{1}{(\hat{\omega}_g)^{\text{new}}(1 + 2\kappa_g/n_g)}\,\text{diag}\{\mathbf{S}_g' - \hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\},$$

and κ_g is as in the CUCU case but with ξ_{gj} given by the jth element along the diagonal of the matrix $\mathbf{S}_g' - \hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g$.

Note that the predicted clustering for each member of the EPGMM family is given by the maximum *a posteriori* (MAP) classification. That is, the posterior predicted component membership of tissue i is the value of g for which $\hat{z}_{ig}$ is greatest.

2.3 Convergence & Model Selection

2.3.1 Convergence Criterion Aitken's acceleration (Aitken, 1926) is used in the analyses herein to estimate the asymptotic maximum of the log-likelihood at each iteration. This allows a decision about whether or not a given AECM algorithm has converged. Aitken's acceleration at iteration t is given by

$$a^{(t)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}},$$

where $l^{(t+1)}$, $l^{(t)}$ and $l^{(t-1)}$ are the log-likelihood values from iterations $t+1$, t and $t-1$, respectively. The asymptotic estimate of the log-likelihood at iteration $t+1$ is given by

$$l_\infty^{(t+1)} = l^{(t)} + \frac{1}{1 - a^{(t)}}(l^{(t+1)} - l^{(t)})$$

(Böhning *et al.*, 1994). Herein, the stopping criterion proposed by McNicholas *et al.* (2010) is used, so that the algorithm can be stopped when $l_\infty^{(t+1)} - l^{(t)} < \epsilon$. More specifically, $\epsilon = 0.1$ is used. Note that this criterion is very similar to that proposed by Lindsay (1995), who suggested stopping when $l_\infty^{(t+1)} - l^{(t+1)} < \epsilon$.

2.3.2 Model Selection The Bayesian information criterion (BIC Schwarz, 1978) is used to select the best member of the EPGMM family, in terms of both model and number of factors. Note that the BIC can also be used to select the number of mixture components (cf. Fraley and Raftery, 1999; McNicholas and Murphy, 2008) but this is not necessary for the analyses herein since we fix $G = 2$. For a model with parameters $\boldsymbol{\theta}$, the Bayesian information criterion (BIC) is given by $\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\theta}}) - m\log n$, where $l(\mathbf{x}, \hat{\theta})$ is the maximized log-likelihood, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, m is the number of free parameters in the model and n is the number of observations. The effectiveness of the BIC for choosing the number of factors in a factor analysis model has been established by Lopes and West (2004), while McNicholas *et al.* (2010) provide practical evidence that the BIC performs well in choosing the number of factors for the PGMM family of models.

A number of other model selection criteria could be used including the Akaike information criterion (AIC Akaike, 1974), the integrated completed likelihood (ICL Biernacki *et al.*, 2000) and clustering stability (cf. von Luxburg, 2009). However, we found that the BIC gave a quick solution and generally good clustering results.

3 ANALYSES

3.1 Dimensionality Reduction

McLachlan *et al.* (2002) analyzed two microarray gene expression data sets — one on leukaemia data and another on colon tissue samples — using the EMMIX-GENE approach. The first stage of this approach focuses on data reduction where, initially, one and two-component mixtures of t-distributions are fitted to the data. Then a gene is retained only if two conditions are satisfied.

One of these conditions is that the minimum cluster size exceeds some pre-specified threshold a_1. The other condition concerns the result of a likelihood ratio test, or tests. First, the hypothesis $H_0 : G = 1$ is tested against $H_1 : G = 2$ and the gene is retained if

$$-2\log\lambda > a_2, \tag{6}$$

where λ is the likelihood ratio statistic. However, if the condition in Equation 6 is not met then the hypothesis $H_0 : G = 2$ is tested against $H_1 : G = 3$ and the gene is retained if the same condition is satisfied, with the same a_2, for this test statistic λ and at least two of the three components contain at least a_1 tissues.

When fitting the two and three-component mixture models for this purpose, starting values for the component memberships are defined randomly or by using starting values based on k-means clustering results. This whole process represents the first stage of the EMMIX-GENE approach and can be carried out using the `select-genes` software that accompanies McLachlan *et al.* (2004). For the analyses herein, the `select-genes` software is used with thresholds $a_1 = a_2 = 8$, as in McLachlan *et al.* (2002), and 50 random and 50 k-means starts.

3.2 Leukaemia Data

3.2.1 The Data Golub *et al.* (1999) presented data on two forms of acute leukaemia: acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). Affymetrix arrays were used to collect measurements for 7,129 genes on 72 tissues. There were a total of 47 ALL tissues and 25 with AML. The data were sourced from the website accompanying McLachlan *et al.* (2004, `www.maths.uq.edu.au/~gjm/emmix-gene/`) and so they had been preprocessed (Dudoit *et al.*, 2002; McLachlan *et al.*, 2002) as follows.

1. Genes with expression less than 100 or greater than 16,000 were removed.
2. Genes with expressions satisfying max/min$\leq$5 and max$-$min$\leq$500 were removed.
3. The natural logarithm was taken.

Following this preprocessing, a total of 3,731 genes remained. This number was further reduced to 2,030 following application of the `select-genes` software (cf. Section 3.1).

3.2.2 The EPGMM Approach Treating this as a clustering problem where the form of leukaemia is unknown, all twelve members of the EPGMM family (Table 2) were fitted to these data for $G = 2$, $q = 1, \ldots, 6$ and ten different random starting values for the $\hat{z}_{ig}$. The BIC for the best q for each of the 12 members of the EPGMM family is given in Table 3.

Table 3. The BIC for the best q for each of the 12 members of the EPGMM family for the leukaemia data.

Model	q	BIC	Model	q	BIC
CCCC	3	-411,646.50	CCUC	3	-411,566.29
UCCC	1	-416,954.56	UCUC	1	-416,803.57
CCCU	4	-414,615.22	CCUU*	5	-413,207.29
UCCU	1	-423,354.79	UCUU*	1	-422,089.38
CUCU*	4	-413,966.90	CUUU	5	-413,978.04
UUCU*	1	-423,933.46	UUUU	1	-423,532.04

* denotes one of the four new models.

The best of these models, in terms of BIC, was a CCUC model with $q = 3$ latent factors. The chosen model has a non-diagonal covariance structure where the covariance between pairs of genes is equal across different clusters but the variance of each gene is unequal across different clusters (see Section 1.2). The MAP classifications arising from the parameter estimates associated with this model are given in Table 4; only five tissue samples were misclassified.

Table 4. Estimated group membership for the best EPGMM model for the leukaemia data.

	1	2
Acute lymphoblastic leukaemia (ALL)	42	0
Acute myeloid leukaemia (AML)	5	25

3.2.3 Hierarchical Clustering, k-means, k-medoids & MCLUST In addition to the EPGMM technique, several other techniques were applied to these data using the R software (R Development Core Team, 2010). Agglomerative hierarchical clustering was used, with Euclidean distance and three different linkage methods: complete, average, and single. The k-means (cf. Hartigan and Wong, 1979) and k-medoids techniques were also used. In the latter case, the partitioning around medoids (PAM; cf. Kaufman and Rousseeuw, 1990, Chapter 2) algorithm was used. Finally, in order to compare our model-based clustering approach to the well-established MCLUST approach, we used the `mclust` package (Fraley and Raftery, 1999) for the R software.

Table 5. Summary results for all of the clustering techniques that were applied to the leukaemia data.

	BIC	Rand Index	Adjusted Rand Index
Hierarchical (Complete)	–	0.532	0.058
Hierarchical (Average)	–	0.525	-0.024
Hierarchical (Single)	–	0.532	-0.013
k-means	–	0.593	0.187
PAM	–	0.518	0.023
MCLUST (VII)	-416,293.2	0.593	0.186
EPGMM (CCUC, $q = 3$)	-411,566.3	0.869	0.738

The results, which are summarized in Table 5, give the Rand and adjusted Rand indices as measures of class agreement.

The Rand index (Rand, 1971) is based on pairwise agreements and disagreements, and the adjusted Rand index (Hubert and Arabie, 1985) is effectively the Rand index corrected for random chance. These indices reveal that the best of the non-model-based approaches was k-means clustering, with an adjusted Rand index of 0.187. In fact, k-means clustering outperformed `mclust` on these data, but the EPGMM model with the greatest BIC (CCUC, $q = 4$) was the best model overall.

3.2.4 The EMMIX-GENE Approach McLachlan *et al.* (2002) analyzed the same data using the EMMIX-GENE approach with four random and four k-means starts in the first stage, which reduced the number of genes to 2,015. In the second stage, a mixture of 40 normal distributions with isotropic covariance structure was fitted to the 2,015 genes. Two of these groups (Groups 1 and 3) provided clusterings that were most similar to the type of luekaemia — of course, in a real clustering scenario this could not be established. A two-component mixture of factor analyzers, with $q = 6$ factors, was fitted to the data using the genes from groups 1 and 3, respectively. Using the genes from Group 1 led to the misclassification of 13 tissues and using those from Group 3 led to the misclassification of 6 tissues. Note that McLachlan *et al.* (2002) did not specify how many different random starts were used but, based on other analyses, it seems likely that 50 random and 50 k-means starts were used.

3.2.5 Two Other Approaches In addition to the EMMIX-Gene approach, McLachlan *et al.* (2002) used two other approaches to cluster the leukaemia tissues. In both cases, the first stage was identical to that described in Section 3.2.4. The first alternative approach was to cluster the tissues based on the 40 fitted group means and the top 50 of the 2,015 genes. Fitting a two-component mixture of factor analyzers with $q = 8$ factors to these data, using 50 random and 50 k-means starts, led to the misclassification of just one tissue. The second alternative approach was to base the analysis on the top fifty genes. Fitting a two-component mixture of factor analyzers, with $q = 8$ factors to these data, using 50 random and 50 k-means starts, led to the misclassification of ten tissues.

3.2.6 Comments The EPGMM approach gave very good clustering performance when applied to the leukaemia data. This approach used ten random starts and led to the misclassification of just five tissues. This performance far exceeds that of agglomerative hierarchical clustering, k-means clustering, PAM, and MCLUST. In fact, the best of all of these techniques had an adjusted Rand index of 0.189, while the best EPGMM model had an adjusted Rand index of 0.738. Although, in one instance, one of the approaches of McLachlan *et al.* (2002) returned a better predicted classification, it is difficult to make a direct comparison to the EPGMM approach. This difficulty arises because the EPGMM approach is a genuine clustering approach, while the methods described in sections 3.2.4 and 3.2.5 assumed, to some extent, knowledge of the truth. This knowledge was clearly used in the analyses described in Section 3.2.4 but was used in a less obvious fashion in the analyses given in Section 3.2.5. In this latter case, the choice of the number of clusters (40) was validated in some sense by the fact that two of the groups give classifications that were similar to the true leukaemia type. In fact, as mentioned by McLachlan *et al.* (2002), an objective technique for choosing this number is not possible since genes cannot be assumed to be independently distributed within a tissue sample. Furthermore, it is quite likely that the number of factors q was selected, in each case, to give the best classification. This could be done objectively, as in Section 3.2.2, using the BIC. Finally, any comparison between the EPGMM approach and the approaches of McLachlan *et al.* (2002) would have to be taken in context with the fact that different subsets of the 3,731 genes are used in each case.

3.3 Colon Data

3.3.1 The Data Alon *et al.* (1999) presented gene expression data on 62 colon tissue samples, of which 40 were tumours and the remaining 22 were normal. Affymetrix arrays were used to collect measurements for 6,500 gene expressions on all 62 tissues. Following Alon *et al.* (1999) and McLachlan *et al.* (2002), only the 2,000 genes with the highest minimal intensity are focused upon. The data were again sourced from the website mentioned in Section 3.2.1 and, this time, the only preprocessing was the taking of natural logarithms, followed by normalization. Application of the `select-genes` software, with the settings specified in Section 3.1, led to the reduction of the number of genes from 2,000 to just 461.

3.3.2 The EPGMM Approach Treating this as a clustering problem where the type of tissue is unknown, all twelve members of the EPGMM family (Table 2) were fitted to these data for $G = 2$, $q = 1, \ldots, 10$ and ten different random starting values for the $\hat{z}_{ig}$. The BIC for the best q for each of the 12 members of the EPGMM family is given in Table 6.

Table 6. The BIC for the best q for each of the 12 members of the EPGMM family for the colon data.

Model	q	BIC	Model	q	BIC
CCCC	4	-79,085.73	CCUC	6	-70,937.72
UCCC	3	-77,267.65	UCUC	3	-77,268.16
CCCU	8	-71,064.11	CCUU*	7	-71,063.71
UCCU	3	-77,310.19	UCUU*	4	-77,532.97
CUCU*	8	-71,609.10	CUUU	8	-71,631.35
UUCU*	2	-78,458.83	UUUU	2	-78,306.33

* denotes one of the four new models.

The best of these models, again in terms of BIC, was a CCUC model with $q = 6$ latent factors; the covariance structure in this model is the same as that chosen for the leukaemia data (see Section 3.2.2). The MAP classifications given by the parameter estimates associated with this model are given in Table 7 ; only five tissue samples were misclassified.

Table 7. Estimated group membership for the best EPGMM model for the colon data.

	1	2
Tumour	37	3
Normal	2	20

3.3.3 Hierarchical Clustering, k-means, k-medoids & MCLUST In addition to the EPGMM technique, the methods used in

Section 3.2.3 were run on these colon data using the R software. The results, which are summarized in Table 8, suggest that the best of the non-model-based approaches was PAM, with an adjusted Rand index of 0.218. This time, `mclust` outperformed k-means clustering but PAM outperformed `mclust`. The EPGMM model with the greatest BIC (CCUC, $q = 6$) was the best model, misclassifying just five tissues based on ten random starts.

Table 8. Summary results for all of the clustering techniques that were applied to the colon data.

	BIC	Rand Index	Adjusted Rand Index
Hierarchical (Complete)	–	0.497	-0.018
Hierarchical (Average)	–	0.526	-0.005
Hierarchical (Single)	–	0.526	-0.014
k-means	–	0.494	-0.016
PAM	–	0.611	0.218
MCLUST (VII)	-81,124.36	0.500	-0.006
EPGMM (CCUC, $q = 6$)	-70,937.72	0.849	0.697

3.3.4 Correspondence with McLachlan et al. *(2002)* Using various techniques, McLachlan *et al.* (2002) found five different clusterings of these data. However, none of these clusterings corresponded to the tissue type. While, once again, the EPGMM results are not directly comparable to those of McLachlan *et al.* (2002), it is interesting to look at the second best of the EPGMM models. The second best of the EPGMM models, in terms of BIC, was a CCUU model with $q = 7$ latent factors. Note that this is one of the four new models that were introduced herein and, again, this model has equal covariance between pairs of genes, however the variance structure is more complex than for the CCUC model. The MAP classification given by the parameter estimates associated with this CCUU model do not separate tumour from normal tissue. However, they are similar to what McLachlan *et al.* (2002) call C_1, in that they seem sensible when one considers that there was a change of protocol during the experiment (Getz *et al.*, 2000; McLachlan *et al.*, 2002). Specifically, tissues 1–11 and 41–51 were all extracted from the first 11 patients using a poly detector, while the remaining samples were taken from the other patients using total extraction of RNA. Looking at the tissues by extraction method, rather than by tissue type, leads to the estimated classifications given in Table 9; only eight of the tissues were misclassified by this CCUU model when the data are considered by extraction method.

Table 9. Estimated group membership for the second best EPGMM model for the colon data.

	1	2
Poly Detector	19	3
Total Extraction of RNA	5	35

The results from applying the other methods to the colon data (cf. Table 8), can also be viewed in terms of extraction method, rather than tissue type. These results are given, along with our best CCUU model, in Table 10. From this table, it is clear that our CCUU model gives the best clustering performance of all of the approaches. Furthermore, the hierarchical (complete and average linkage), k-means, and MCLUST clustering results are all better when viewed in terms of extraction method.

Table 10. Summary results, by extraction method, for all of the clustering techniques that were applied to the colon data.

	BIC	Rand Index	Adjusted Rand Index
Hierarchical (Complete)	–	0.518	0.024
Hierarchical (Average)	–	0.545	0.035
Hierarchical (Single)	–	0.526	-0.014
k-means	–	0.526	0.048
PAM	–	0.581	0.158
MCLUST (VII)	-81,124.36	0.526	0.045
EPGMM (CCUU, $q = 7$)	-71,063.71	0.772	0.542

3.3.5 Comments The EPGMM approach gave very good clustering performance when applied to the colon data. Our approach led to the misclassification of just five tissues, when these data were viewed by tissue type. This performance far exceeded that of all of the other techniques that were used — in fact, the performance of these other approaches was surprisingly poor, with only PAM giving better than random classifications (cf. Table 8). This phenomenon is partly explained when one looks at the classifications by extraction method, rather than by tissue (cf. Table 10). In this case, only one method performed worse than random, which might suggest that techniques like k-means clustering and MCLUST were picking up extraction method more-so than tissue type. That said, the performance of these methods was only slightly better than random which suggests that the restrictive cluster shapes imposed by k-means clustering and MCLUST were not at all suited to the data. On the other hand, the best of the new EPGMM models gave the based clustering performance, misclassifying just eight samples.

4 DISCUSSION

The EPGMM family of models has been shown to give good clustering performance when applied to gene expression microarray data. These applications, concerning leukaemia and colon tissue data, respectively, were conducted as genuine clustering examples. That is, no information on the true tissue classification was used for parameter estimation or model selection. In fact, this information was only used to assess the performance of the selected model. In this context, the clustering performance of the EPGMM family can be looked upon favourably. Moreover, the performance of the EPGMM family on both data sets far exceeded that of a number of popular clustering techniques, including agglomerative hierarchical clustering and k-means clustering.

However, like the techniques of McLachlan *et al.* (2002), the EPGMM family relies on multiple random starts. In addition to the obvious drawback of the sensitivity of results to the starting values, there is the computation time that is required. Furthermore, there is no guarantee that increasing the number of random starts will lead to better clustering results. This is due, in the main, to the fact that models with greater BIC do not necessarily give better clustering performance. This phenomenon has been observed previously and

work into finding better model selection techniques is ongoing. That said, the EPGMM family did perform well in the analyses in Section 3, based on random starting values and using the BIC.

5 CONCLUSION

The EPGMM family of mixture models, for the model-based clustering of gene expression microarray data, has been introduced. This family of models is an extension of the PGMM family of models which, in turn, is an extension of the mixtures of factor analyzers model. The EPGMM family of models are very well suited to the analysis of high dimensional data. The reason for this suitability is three-fold. First, each member of the EPGMM family has a number of covariance parameters that is linear in the data-dimensionality. Second, as shown herein, the Woodbury identity can be used to avoid the inversion of any non-diagonal $p \times p$ matrices, leading to efficient computation. Thirdly, as shown by McNicholas *et al.* (2010) in the context of the PGMM family, these models are 'trivially parallelizable', opening up the possibility of even more efficient parameter estimation using parallel computing. The EPGMM family was applied to two well known gene expression microarray data sets. In both cases, the EPGMM family performed well and gave much better clusterings than several popular clustering techniques. Herein, we took $G = 2$ for all of the analysis but future work will focus on the selection of G.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the insightful and helpful comments of three anonymous reviewers. This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, and by a Science Foundation Ireland Basic Research Grant (04/BR/M0057). The high-performance computing equipment that was used was provided by Silicon Graphics Inc. through funding from the Canada Foundation for Innovation–Leaders Opportunity Fund and from the Ontario Research Fund–Research Infrastructure Program.

REFERENCES

Aitken, A. C. (1926). On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, **46**, 289–305.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(12), 6745–6750.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725.

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*, **46**, 373–388.

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

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**(457), 77–87.

Fraley, C. and Raftery, A. E. (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification*, **16**, 297–306.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York.

Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(22), 12079–12084.

Ghahramani, Z. and Hinton, G. E. (1997). The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University Of Toronto, Toronto.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*, **28**, 100–108.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley, New York.

Lagrange, J. L. (1788). *Méchanique Analitique*. Chez le Veuve Desaint, Paris.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, volume 5. Hayward, California: Institute of Mathematical Statistics.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**, 41–67.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley, New York, 2nd edition.

McLachlan, G. J. and Peel, D. (2000a). *Finite Mixture Models*. John Wiley & Sons, New York.

McLachlan, G. J. and Peel, D. (2000b). Mixtures of factor analyzers. In P. Langley, editor, *Seventh International Conference on Machine Learning*, pages 599–606, San Francisco. Morgan Kaufmann.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**(3), 412–422.

McLachlan, G. J., Do, K.-A., and Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Wiley, Hoboken, New Jersey.

McLachlan, G. J., Bean, R. W., and Jones, L. B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**(13), 1608–1615.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. *The Canadian Journal of Statistics*, **38**(1), 153–168.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**(3), 711–723.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Meng, X. L. and van Dyk (1997). The EM algorithm — an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society Series B*, **59**, 511–567.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 31–38.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, **15**, 72–101.

Tipping, T. E. and Bishop, C. M. (1999). Mixtures of probabilistic principal component analysers. *Neural Computation*, **11**(2), 443–482.

von Luxburg, U. (2009). Clustering stability: An overview. *Foundations and Trends in Machine Learning*, **2**(3), 235–274.

Woodbury, M. A. (1950). *Inverting modified matrices*. Statistical Research Group, Memo. Rep. no. 42. Princeton University, Princeton, New Jersey.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A., and Ruzzo, L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.