BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 14 2006, pages 1745–1752
doi:10.1093/bioinformatics/btl165

*Gene expression*

# A Mixture model with random-effects components for clustering correlated gene-expression profiles

S. K. Ng[1], G. J. McLachlan[1,2,3,*], K. Wang[3], L. Ben-Tovim Jones[2] and S.-W. Ng[4]

[1]Department of Mathematics, [2]Institute for Molecular Bioscience and [3]ARC Centre for Complex Systems, University of Queensland, Brisbane, QLD 4072, Australia and [4]Laboratory of Gynecologic Oncology, Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Boston, MA 02115, USA

Received on February 15, 2006; revised on April 10, 2006; accepted on April 26, 2006
Advance Access publication May 3, 2006
Associate Editor: Martin Bishop

## ABSTRACT

**Motivation:** The clustering of gene profiles across some experimental conditions of interest contributes significantly to the elucidation of unknown gene function, the validation of gene discoveries and the interpretation of biological processes. However, this clustering problem is not straightforward as the profiles of the genes are not all independently distributed and the expression levels may have been obtained from an experimental design involving replicated arrays. Ignoring the dependence between the gene profiles and the structure of the replicated data can result in important sources of variability in the experiments being overlooked in the analysis, with the consequent possibility of misleading inferences being made. We propose a random-effects model that provides a unified approach to the clustering of genes with correlated expression levels measured in a wide variety of experimental situations. Our model is an extension of the normal mixture model to account for the correlations between the gene profiles and to enable covariate information to be incorporated into the clustering process. Hence the model is applicable to longitudinal studies with or without replication, for example, time-course experiments by using time as a covariate, and to cross-sectional experiments by using categorical covariates to represent the different experimental classes.

**Results:** We show that our random-effects model can be fitted by maximum likelihood via the EM algorithm for which the E(expectation) and M(maximization) steps can be implemented in closed form. Hence our model can be fitted deterministically without the need for time-consuming Monte Carlo approximations. The effectiveness of our model-based procedure for the clustering of correlated gene profiles is demonstrated on three real datasets, representing typical microarray experimental designs, covering time-course, repeated-measurement and cross-sectional data. In these examples, relevant clusters of the genes are obtained, which are supported by existing gene-function annotation. A synthetic dataset is considered too.

**Availability:** A Fortran program blue called EMMIX-WIRE (EM-based MIXture analysis WIth Random Effects) is available on request from the corresponding author.

**Contact:** gjm@maths.uq.edu.au

**Supplementary information:** http://www.maths.uq.edu.au/~gjm/bioinf0602_supp.pdf. Colour versions of Figures 1 and 2 are available as Supplementary material on *Bioinformatics* online.

## 1 INTRODUCTION

In recent times, (mixture) model-based clustering has become very popular in the cluster analysis of microarray data (Ghosh, and Chinnaiyan, 2002; Yeung *et al*., 2001; McLachlan *et al*., 2002; Medvedovic and Sivaganesan, 2002; among others), as it provides a sound mathematical framework for clustering; see McLachlan, and Basford (1988), Fraley and Raftery (1998), and McLachlan and Peel (2000). With this approach to clustering, a common assumption in practice is to take the component densities to be multivariate normal, which is computationally tractable and ensures that the implied clustering is invariant under changes in location and scale, as well as rotation, of the data. However, in unmodified form, this approach does not incorporate experimental design information such as disease status of the tissue samples in which the genes are measured in cross-sectional studies, covariate information such as the time ordering of the gene measurements in time-course studies or the structure of the replicated data as in longitudinal studies.

Recently, Pan (2006) has proposed to incorporate known gene functions as prior probabilities in model-based clustering. blue Previously, Cheng *et al*. (2004) had used the graph-based structure of Gene Ontology (GO) for inferring the similarity between genes. But there is a need to develop further clustering procedures that are applicable to data from a wide variety of experimental designs blue that can be applied without the aid of biological knowledge. This is because present databases are necessarily incomplete and evolving. It is hoped that the clustering of the genes will reveal new biological knowledge that in time will be represented in the annotation schemes (Clare and King, 2002; Gibbons and Roth, 2002). There is also the need to develop further clustering procedures that are applicable to data from a wide variety of experimental designs. For example, microarray experiments are now being carried out with replication for capturing either biological or technical variability in expression levels to improve the quality of inferences made from experimental studies (Lee *et al*., 2000; Pavlidis *et al*., 2003). Replicated measurements from each tissue sample (subject) are often interdependent and tend to be more alike in characteristics than data chosen at random from the population as a whole. Similarly, in time-course studies (Storey *et al*., 2005) where expression levels are measured under various conditions or at different time points, gene expressions obtained from the same condition (subject)

[*]To whom correspondence should be addressed.

are correlated. Ignoring the dependence between microarray data can lead to misleading inferences being made (Luan and Li, 2003).

In this paper, we consider the extension of normal mixture models to correlated and replicated data via the formulation of a (multilevel) linear mixed-effects model (LMM) for the mixture components in which covariate information can be incorporated. With this LMM, subject effects are assumed to be random (random effects) and shared among expression levels obtained from the same subject (McCulloch and Searle, 2001). Our contribution is to create a general framework for the mixture model-based clustering of correlated genes, based on expression microarray data obtained from various experimental designs such as repeated measurement data and time-course data. The proposed general random-effects model framework is formulated by incorporating both 'gene' effects and 'tissue' effects in the mixture modeling of the microarray data. This is in contrast to the mixed-effects models approaches in Celeux *et al.* (2005), Luan and Li (2003) and McLachlan *et al.* (2004) that involve only gene-specific random effects. Their methods thus require the independence assumption for the genes which, however, will not hold in practice for all pairs of genes (McLachlan *et al.*, 2004; Klebanov *et al.*, 2006). In the context of modeling somatic cell counts in dairy cattle, Gianola *et al.* (2004) and Ødegård *et al.* (2005) have proposed a finite mixture of mixed models for univariate data, in which dependence among the observations is induced by taking the covariance matrix in the distribution of the individual random effects terms to be non-diagonal. However, it is taken to be known (determined by the pedigree structure), whereas we infer the correlations via estimation of the variances of the random effects by fitting our postulated (multivariate) model to the data. And we implement the EM algorithm exactly without the need to use Monte Carlo methods to carry out the E-step as in Gianola *et al.* (2004).

The proposed framework of LMMs is not limited to the clustering of genes, which is the focus of this paper. It can also be readily adopted to detect differentially expressed genes in known classes of tissue samples based on a normal mixture model approach (Lee *et al.*, 2000; Pan *et al.*, 2002) and as a gene-reduction method in the classification of tissue samples (McLachlan *et al.* 2002). The mixture framework of LMMs approach is to be illustrated in the clustering of three representative datasets in the microarray literature, namely the yeast cell-cycle data of Spellman *et al.* (1998), the yeast galactose data of Ideker *et al.* (2001), and the colorectal carcinoma data of Muro *et al.* (2003). A synthetic dataset is also considered.

## 2 LINEAR MIXED-EFFECTS MODELS

Although biological experiments vary considerably in their design, the data generated by microarray experiments can be viewed as a matrix of expression levels. For $m$ microarray experiments (corresponding to various tissue samples, tissue types, repeated measurements or time points), where we measure the expression levels of $n$ genes in each experiment, the microarray data can be represented by an $n \times m$ matrix. A general framework of random-effects model is formulated by incorporating both gene effects and tissue effects in the mixture modeling of the microarray data. We let $\boldsymbol{y}_j = (\boldsymbol{y}_{1j}, \ldots, \boldsymbol{y}_{mj})^{\mathrm{T}}$ denote the measurement on the $j$-th gene, where the superscript T denotes vector transpose. It is assumed that the expression levels have been preprocessed with adjustment for array effects.

With the mixture model-based clustering approach, the observed $m$ -dimensional vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ 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 $h$-th component of the mixture, the vector $y_j$ for the $j$-th gene follows the model

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta}_h + \boldsymbol{U}\boldsymbol{b}_{hj} + \boldsymbol{V}\boldsymbol{c}_h + \boldsymbol{\epsilon}_{hj}, \qquad (1)$$

where elements of $\boldsymbol{\beta}_h$ (a $p$-dimensional vector) are fixed effects (unknown constants) modeling the conditional mean of $y_j$ in the $h$-th component, $\boldsymbol{b}_{hj}$ (a $q_b$-dimensional vector) and $\boldsymbol{c}_h$ (a $q_c$-dimensional vector) represent the unobservable random gene effects and tissue effects, respectively. The random effects $\boldsymbol{b}_{hj}$ and $\boldsymbol{c}_h$, and the measurement error vector $\boldsymbol{\epsilon}_{hj}$ are assumed to be mutually independent. In (1), $\boldsymbol{X}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ are known design matrices of the corresponding fixed or random effects. The specification of (1) covers many general random-effects models for the clustering of correlated gene expression data arisen from various microarray experiments, including those with replications. For example, let $t$ be the number of distinct tissues in the experiment. We are given for the $j$-th gene a feature vector $\boldsymbol{y}_j = (\boldsymbol{y}_{1j}^T, \ldots, \boldsymbol{y}_{tj}^T)^T$, where $\boldsymbol{y}_{lj} = (y_{l1j}, \ldots, y_{lrj})^T$ contains the $r$ replications on the $j$-th gene from the $l$-th tissue ($l = 1, \ldots, t$). With respect to (1), $\boldsymbol{\beta}_h$ is a $p$-dimensional vector ($p = t$) modeling the conditional mean of $\boldsymbol{y}_j$ in the $h$-th component. Moreover, conditional on membership of the $h$-th component, it is assumed that the random effects are shared among the repeated measurements of expression on the same gene from the same tissue [$\boldsymbol{b}_{hj}$ in (1) with $q_b = t$], along with the random effects that are shared among gene expressions from the same tissue [$\boldsymbol{c}_h$ in (1) with $q_c = m = tr$]. The component-specific effects $\boldsymbol{c}_h$ for the tissues induce dependency among the gene-expression levels of genes from the same component and from the same tissue (correlated genes). By allowing the expression levels of genes in the same cluster to be correlated, blue the genes in a cluster to have their own and cluster-specific random-effects terms, there can be greater individual and collective variation, respectively, exhibited by the genes in the same cluster than otherwise possible under a fixed-effects model without gene- and cluster-specific random effects.

With the LMM, the distributions of $\boldsymbol{b}_{hj}$ and $\boldsymbol{c}_h$ are taken to be multivariate normal, $N_{q_b}(\mathbf{0},\theta_{bh} \quad \mathbf{I}_{\mathrm{q}_b})$ and $N_{q_c}(\mathbf{0},\theta_{ch} \quad \boldsymbol{I}_{q_c})$, respectively, where $\boldsymbol{I}_{qb}$ and $\boldsymbol{I}_{\mathrm{qc}}$ are identity matrices with dimensions being specified by blue their subscripts. The measurement error vector $\boldsymbol{\epsilon}_{hj}$ is also taken to be multivariate normal $N_m(\mathbf{0}, \boldsymbol{A}_{\mathrm{h}})$, where $\boldsymbol{A}_h = \mathrm{diag}\ (\boldsymbol{W}\phi_h)$ is a diagonal matrix constructed from the vector $(\boldsymbol{W}\phi_h)$ with $\phi_h = (\sigma_{h1}^2, \ldots, \sigma_{hq_e}^2)^{\mathrm{T}}$ and $\mathbf{W}$ a known $m \times q_e$ zero-one design matrix. That is, we allow the $h$-th component-variance to be different among the $m$ microarray experiments.

Genes from the same component (i.e. within the same cluster) have a common term ($\boldsymbol{V}\boldsymbol{c}_{\mathbf{h}}$ in the case of the $h$-th cluster) in the linear model (1) for their expression levels. As this term is a random rather than a fixed effect one, it means that genes within the same cluster are correlated.

## 3 ML ESTIMATION VIA THE EM ALGORITHM

We let $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1^T, \ldots, \boldsymbol{\psi}_g^T, \pi_1, \ldots, \pi_{g-1})^T$ be the vector of all the unknown parameters, where $\boldsymbol{\psi}_h$ is the vector containing the unknown parameters $\boldsymbol{\beta}_h$, $\theta_{bh}$, $\theta_{ch}$ and $\boldsymbol{\phi}_h$ of the $h$-th component density ($h = 1, \ldots, g$). The estimation of $\Psi$ can be obtained by maximum likelihood (ML) via the EM algorithm (Dempster *et al.*, 1977). The implementation of the E-step is straightforward for mixture models provided that the data can be treated as being independently distributed. In our model (1), the gene-profile vectors $\boldsymbol{y}_j$ are not all independently distributed as genes within the same cluster (i.e. from the same component in the mixture model) are allowed to be dependent due to the presence of the random-effects term $\boldsymbol{c}_h$ for the $h$-th component in (1). However, we can circumvent this problem by proceeding conditionally on the random-cluster effects $\boldsymbol{c}_h$, as given these terms, the gene profile vectors $\boldsymbol{y}_j$ are all conditionally independent. In this way, we can actually carry out the E- and M-steps in closed form. In particular, we do not have to approximate the E-step by carrying out time-consuming Monte Carlo approximations.

Within the EM framework, each $\boldsymbol{y}_j$ is conceptualized to have arisen from one of the $g$ components. We let $z_1, \ldots, z_n$ 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 $\boldsymbol{y}_j$ does or does not come from the $h$-th component given $\boldsymbol{c}$, where $\boldsymbol{c} = (\boldsymbol{c}_1^T, \ldots, \boldsymbol{c}_g^T)^T$. We let $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ \ldots, \boldsymbol{y}_n^T)^T$ denote the observed data and, correspondingly, put $\boldsymbol{z}^T = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)$. The ML estimation of the normal mixture of LMMs via the EM algorithm can be formulated by treating the unobservable component-indicator variables $z$ and the random effects, $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_g^T)^T$ and $\boldsymbol{c}$, as missing data in the EM framework (Ng *et al.*, 2004), where $\boldsymbol{b}_h = (\boldsymbol{b}_{h1}^{\mathrm{T}}, \ldots, \boldsymbol{b}_{hn}^{\mathrm{T}})^{\mathrm{T}}$ for $h = 1, \ldots, g$. By assuming that the random effects are normally distributed, it follows from standard normal theory that the joint distribution of the complete data $(\boldsymbol{y}^T, \boldsymbol{z}^T, \boldsymbol{b}^T, \boldsymbol{c}^T)^T$ is also a normal mixture. Let $\boldsymbol{\epsilon}_h = (\boldsymbol{\epsilon}_{h1}^T, \ldots, \boldsymbol{\epsilon}_{hn}^T)$ for $h = 1, \ldots, g$. The complete data are then given by $(\boldsymbol{y}^{\mathrm{T}}, \boldsymbol{z}^{\mathrm{T}}, \boldsymbol{b}^{\mathrm{T}}, \boldsymbol{c}^{\mathrm{T}})^{\mathrm{T}}$. As the observed data $y$ and the gene-specific random effects $\boldsymbol{b}$ are jointly normal, conditional on $\boldsymbol{z}$ and the cluster-specific random effects $\boldsymbol{c}$, it follows (see Supplementary information) that the complete-data log likelihood is given, apart from an additive constant, by

$$\begin{aligned}\log L_c(\boldsymbol{\Psi}) = \sum_{h=1}^{g}\Bigg[ & \sum_{j=1}^{n} z_{hj}\log\pi_h \\ & - \frac{1}{2}\Bigg\{\sum_{j=1}^{n} z_{hj}q_b\log\theta_{bh} + q_c\log\theta_{ch} \\ & + \sum_{j=1}^{n} z_{hj}\log|\boldsymbol{A}_h| + \frac{\boldsymbol{b}_h^T\boldsymbol{b}_h}{\theta_{bh}} + \frac{\boldsymbol{c}_h^T c_h}{\theta_{ch}} + \boldsymbol{\epsilon}_h^T\Omega_h\boldsymbol{\epsilon}_h\Bigg\}\Bigg],\end{aligned} \tag{2}$$

where

$$\boldsymbol{\epsilon}_h = (\boldsymbol{\epsilon}_{h1}^T, \ \ldots, \boldsymbol{\epsilon}_{hn}^T)^T, \tag{3}$$

$$\boldsymbol{b}_h^{\mathrm{T}}\boldsymbol{b}_h = \sum_{j=1}^{n} z_{hj}\boldsymbol{b}_{hj}^{\mathrm{T}}\boldsymbol{b}_{hj}, \tag{4}$$

and

$$\Omega_h = \boldsymbol{I}_n \otimes \boldsymbol{A}_h^{-1} \tag{5}$$

for $h = 1, \ldots, g$, and hence

$$\boldsymbol{\epsilon}_h^{\mathrm{T}}\Omega_h\boldsymbol{\epsilon}_h = \sum_{j=1}^{n} z_{hj}\boldsymbol{\epsilon}_{hj}^T\boldsymbol{A}_h^{-1}\boldsymbol{\epsilon}_{hj}. \tag{6}$$

In the above, the sign $\otimes$ denotes the Kronecker product of two matrices.

The EM algorithm proceeds by alternating the E- and M-steps where, on the $(k+1)$-th iteration, the E-step involves the calculation of the $Q$-function which is the expectation of the complete-data log likelihood over the joint distribution of the unobservable data $(\boldsymbol{z}^T, \boldsymbol{b}^T, \boldsymbol{c}^T)^T$ given the observed data $y$, using the current estimate $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. It follows from (2) that the E-step involves the calculation of the following conditional expectations,

$$E_{\boldsymbol{\Psi}^{(k)}}(z_{hj} \mid \boldsymbol{y}), \quad E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{b}_h \mid \boldsymbol{y}), \quad E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{b}_h^T\boldsymbol{b}_h \mid \boldsymbol{y}),$$

$$E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{c}_h \mid \boldsymbol{y}), \quad E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{c}_h^T\boldsymbol{c}_h \mid \boldsymbol{y}), \quad E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\epsilon}_h^T\Omega_h\boldsymbol{\epsilon}_h \mid \boldsymbol{y}).$$

These conditional expectations are directly obtainable as shown in the Supplementary material.

The M-step provides the updated estimate $\boldsymbol{\Psi}^{(k+1)}$ that maximizes the $Q$-function with respect to $\boldsymbol{\Psi}$. With reference to (2), the updating formulae for $\boldsymbol{\Psi}^{(k+1)}$ exist in closed form. The detailed derivation is provided in the Supplementary material. The E- and M-steps are alternated repeatedly until convergence of the EM sequence of iterates (Ng *et al.*, 2004).

To effect a probabilistic or an outright clustering of the genes into $g$ components, we condition on the cluster random-effects vector $\boldsymbol{c}_h$. As the latter is unobservable, we use its estimated conditional expectation given the observed data,

$$\hat{\boldsymbol{c}}_h = E_{\hat{\boldsymbol{\Psi}}}(\boldsymbol{c}_h \mid \boldsymbol{y}), \tag{7}$$

where $E_{\hat{\boldsymbol{\Psi}}}$ denotes taking expectation using the ML estimate $\hat{\boldsymbol{\Psi}}$ for the vector $\boldsymbol{\Psi}$ of unknown parameters. Since the genes within a cluster are independently distributed given $\boldsymbol{c}_h$, it suffices to effect a clustering with each gene considered individually in terms of its estimated posterior probabilities of component membership given its profile vector and $\boldsymbol{c}_h$, for $h = 1, \ldots, g$ and $j = 1, \ldots, n$. Using Bayes' theorem, the posterior probability that the $j$-th gene belongs to the $h$-th component given $\boldsymbol{y}_j$ and $\boldsymbol{c}, \tau_h(\boldsymbol{y}_j, \boldsymbol{c}; \boldsymbol{\Psi})$, can be expressed as

$$\begin{aligned}\tau_h(\boldsymbol{y}_j, \boldsymbol{c}; \boldsymbol{\Psi}) &= \mathrm{pr}\{Z_{hj} = 1 \mid \boldsymbol{y}_j, \boldsymbol{c}\} \\ &= \frac{\pi_h f(\boldsymbol{y}_j \mid z_{hj} = 1, \boldsymbol{c}_h; \boldsymbol{\psi}_h)}{\sum_{i=1}^{g} \pi_i f(\boldsymbol{y}_j \mid z_{ij} = 1, \boldsymbol{c}_i; \boldsymbol{\psi}_i)},\end{aligned} \tag{8}$$

where $f(\boldsymbol{y}_j \mid z_{hj} = 1, \boldsymbol{c}_h; \boldsymbol{\psi}_h)$denotes the $h$-th component density of $\boldsymbol{y}_j$ given the random effect $\boldsymbol{c}_h$. The log of this density is given, apart from an additive constant, by

$$\begin{aligned}&\log\ f(\boldsymbol{y}_j \mid z_{hj} = 1, \boldsymbol{c}_h; \boldsymbol{\psi}_h) \\ &= -\frac{1}{2}\left\{\log|\boldsymbol{B}_h| + (\boldsymbol{y}_j - \mathbf{X}\boldsymbol{\beta}_h - \boldsymbol{V}\boldsymbol{c}_h)^{\mathrm{T}}\boldsymbol{B}_h^{-1}(\boldsymbol{y}_j - \boldsymbol{X}\boldsymbol{\beta}_h - V\boldsymbol{c}_h)\right\},\end{aligned} \tag{9}$$

which apart from the additive constant is the log of the $h$-th component density of $\boldsymbol{y}_j$ conditional on $\boldsymbol{c}_h$, where $\boldsymbol{B}_h = \boldsymbol{A}_h + \theta_{bh}\ \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$. Conditional on the cluster random-effects vector $\boldsymbol{c}$, the posterior probabilities (8) of component membership can be used to define the optimal or Bayes rule of allocation for assigning the $j$-th gene with profile vector $\boldsymbol{y}_j$ to one of the $g$ components of the mixture mdoel; see, for example, McLachlan (1992, Chapter 1). The $j$-th gene is assigned outright to the component for which $\tau_h(\boldsymbol{y}_j, \boldsymbol{c}; \boldsymbol{\Psi})$ is greatest for $h = 1, \ldots, g$; that is, the $j$-th gene is assigned outright to component $h^*$, where

$$h^* = \arg \max_{h} \tau_h(\boldsymbol{y}_j, \boldsymbol{c}; \boldsymbol{\Psi}). \tag{10}$$

Here we do not know $\boldsymbol{c}$ nor do we know the parameter vector $\boldsymbol{\Psi}$. Thus we assign the genes on the basis of (10), using $\hat{\boldsymbol{c}}_h$ for $\boldsymbol{c}_h$ and $\hat{\boldsymbol{\Psi}}$ for $\boldsymbol{\Psi}$.

## 4 MODEL SELECTION

The specification of the random-effects components in model (1) needs careful consideration. An identifiability problem could arise if the random-effects model is so specified such that the design matrix $\mathbf{V}$ for the random effect $\boldsymbol{c}_h$ is the same as the $\boldsymbol{X}$ for the fixed effects $\boldsymbol{\beta}_h$. As described in Section 1, many kinds of microarray data have a hierarchical structure. Such data hierarchies may be present naturally or may be due to the experimental design (Goldstein, 1995).

In this study, the emphasis is on the grouping of the genes rather than on the number of clusters and their link with externally existing groups. That is, we are interested primarily in which genes are put together in the same cluster for plausible choices of the number of components $g$ in the mixture model. A guide to plausible values of $g$ can be obtained using BIC (the Bayesian information criterion) of Schwarz (1978), whereby the number $g$ of components in the mixture model is taken to minimize $-2\log L(\hat{\boldsymbol{\Psi}}) + d\log n$, and $d$ denotes the number of parameters in the model. In the EM framework, $L(\boldsymbol{\Psi})$ is the incomplete-data likelihood function for $\boldsymbol{\Psi}$. However, as the gene-profile vectors $\boldsymbol{y}_j$ are not all independently distributed, we are unable to calculate this likelihood function $L(\boldsymbol{\Psi})$ directly by taking the product of the (marginal) densities of the $\boldsymbol{y}_j$. Here we approximate $L(\boldsymbol{\Psi})$ by forming it as if all the $\boldsymbol{y}_j$ were independent. Another approach would be to use resampling methods (Efron and Tibshirani, 1993; McLachlan 1987; McLachlan and Khan, 2004).

## 5 EXAMPLES

We illustrate our method by applying it to three representative datasets, each arising from different kinds of microarray experiments: time course data as in the yeast cell-cycle study of Spellman *et al.* (1998), data with repeated measurements as in the yeast galactose study of Ideker *et al.* (2001) and finally cross-sectional data involving two groups of tissues (tumor and normal) as in the study of human colorectal carcinomas of Muro *et al.* (2003). This section thus demonstrates how the proposed method can be applied to correlated gene-expression array data collected under various experimental designs. We also consider a synthetic dataset with a time-course structure based on our model as fitted to the aforementioned yeast cell-cycle data of Spellman *et al.* (1998).

### 5.1 Clustering of time-course data

By analyzing cDNA microarrays from yeast cultures synchronized by three independent methods over approximately two cell-cycle periods, Spellman *et al.* (1998) identified 800 yeast genes that meet an objective minimum criterion for cell cycle regulation. In our study, we consider the 18 $\alpha$-factor (pheromone) synchronization where the yeast cells were sampled at 7 min intervals for 119 mins. We worked with a subset of 612 genes that have no missing expression data across any of the 18 time points (Luan and Li, (2003).

Our aim here is to cluster the cell cycle-regulated genes based on the microarray expression data matrix of $n = 612$ rows (genes) and $m = 18$ columns (time points). We then analyzed the clusters so formed for common regulatory elements, as described by Spellman *et al.* (1998). With reference to (1), we take the design matrix $\mathbf{X}$ to be an $18 \times 2$ matrix with the $(l + 1)$-th row $(l = 0, \ldots, 17)$

$$(cos(2\pi(7l)/\omega + \Phi) \ldots sin(2\pi(7l)/\omega + \Phi)), \tag{11}$$

where $\omega$ is the period of the cell cycle and $\Phi$ is the phase offset (Spellman *et al.*, 1998). We adopted here the least squares estimation approach considered by Booth *et al.* (2004) to obtain the cell cycle period $\omega = 53$ and the initial phase $\Phi = 0$ from the dataset. For the design matrices of the random effects parts, we take $\boldsymbol{U} = \mathbf{1}_{18}$ and $\boldsymbol{V} = \mathbf{I}_{18}$. That is, we assume that there exists random gene effects $b_{hj}$ with $q_b = 1$ and random temporal effects $(c_{h1}, \ldots, c_{hq_c})$ with $q_c = m = 18$. The latter introduce interdependency among expression levels within the same cluster obtained from the same time point. Also, we take $\boldsymbol{W} = \mathbf{1}_{18}$ and $\boldsymbol{\phi}_h = \sigma_h^2$ $(q_e = 1)$ so that the component variances are common among the $m = 18$ experiments. The mixture model of LMMs as described in Section 2 was fitted to the data with $g = 4$ to $g = 15$ components. The number of components $g$ was determined using BIC for model selection. It indicated here that there are twelve clusters. The clustering results for $g = 12$ are given in Figure 1, where the expression profiles for genes in each cluster are presented. From Figure 1, it can be seen that the genes have very similar expression patterns within each cluster, except in clusters 4 and 7, where there is greater individual variation by some of the genes. This clustering result is different from Spellman's clustering, which was based on time of peak expression only (Spellman *et al.*, 1998). It can be seen from Figure 1 that the genes have very similar expression patterns within each cluster, except in clusters 4 and 7, where there is greater individual variation by some of the genes. This individual variation is permissible under our model which, from the perspective of parsimony, has gene- and cluster-specific random-effects terms to allow for greater variation by the genes from their cluster means than otherwise possible with fixed-effects models.

For Clusters 1, 3, 10, 11 and 12 that show clear periodic expression patterns, we searched through the 700 bp upstream region of the start codon of each gene for the presence of binding site sequences for any known yeast cell cycle transcription factors like MBF, SBF, Mcm1p-containing factors and Swi5p factors. The results are summarized in Table 1. We found that the majority of the genes in these clusters share common promoter elements, and furthermore, they correspond to known cell-cycle transcription factor binding sites relevant to the time of peak expression. For example, genes in Clusters 1 and 10 show typical G1 peak expression and were the major members of the 'CLN2' cluster described by Spellman *et al.* (1998). But there is a higher percentage (45%) of genes in Cluster 10 that also contain a Swi5p site compared with the 28% of genes in Cluster 1. This supports our findings that the 'CLN2' gene cluster corresponds to two distinct groups and that these may be under different regulatory control. Genes in Cluster 3 contain genes previously clustered in the 'CLB2' cluster of Spellman's work. These genes include CLB1, CDC5 and CDC20, which are involved in mitosis and peak in the M phase. Clusters 11 and 12 of our cluster analysis contain, respectively, members of the 'Y' cluster and histone genes described by Spellman *et al.* (1998). The expression of histone genes is tightly peaked in

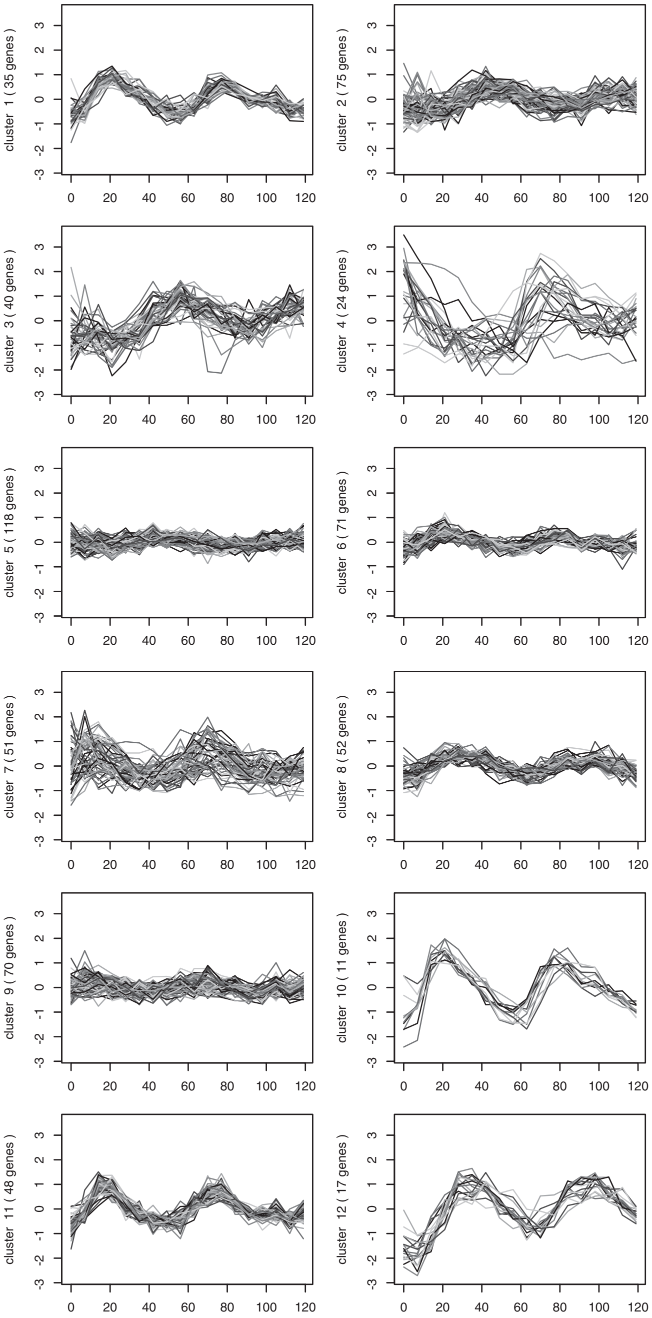

**Fig. 1.** Clustering results for the yeast cell-cycle data. For all the plots, the *X*-axis is the time point and the *Y*-axis is the gene-expression level. A colour version of this figure is available as Supplementary material.

**Table 1.** Promoter elements (Yeast cell-cycle data)

| Cluster | No. of genes | Binding site | Regulator | Peak expression |
|---|---|---|---|---|
| 1 | 35 | ACGCGT | MBF, SBF | $G_1$ |
| 3 | 40 | MCM1 + SFF | Mcm1p + SFF | $G_2$/M |
| 10 | 11 | ACGCGT | MBF, SBF | $G_1$ |
| 11 | 48 | Unknown | Unknown | $G_1$ |
| 12 | 17 | ATGCGAAR | Unknown | S |

the S phase, and there are very high peak-to-trough ratios. The detailed description of the results of the analysis of common regulatory elements is given in the Supplementary information.

## 5.2 Clustering of genes with repeated measurements

This dataset has been used to study integrated genomic and proteomic analyses of a systemically perturbed metabolic network (Ideker *et al.*, 2001) and is available from the online version of Yeung *et al.* (2003). With this yeast galactose data, there are four ($r = 4$) replicate hybridizations for each cDNA array experiment. However, ~8% of the data are missing. A *k*-nearest neighbor ($k = 12$) method has been adopted to impute all the missing values (Yeung *et al.*, 2003). There are $n = 205$ genes and $t = 20$ tissues. The expression patterns of these 205 genes reflect four functional categories in the Gene Ontology (GO) listings (Ashburner *et al.*, (2000); Yeung *et al.*, (2003). In our study, we take $m = tr = 80$ and $\boldsymbol{X} = \mathbf{1}_4 \otimes \boldsymbol{I}_{20}$ (a 80 × 20 matrix). The design matrix $\boldsymbol{U}$ is taken to be equal to $\boldsymbol{X}$ ($q_b = 20$) and $\boldsymbol{V}$ is taken to be $\boldsymbol{I}_{80}$ ($q_c = m = 80$). That is, we assumed that there exists random effects that are shared among the repeated measurements of expression on the same gene from the same tissue [$\mathbf{b}_{hj}$ in (1)]. At the same time, there exists random effects that are shared among gene expressions from the same tissue [$\boldsymbol{c}_h$ in (1)]. In this study, we allow the *h*-th component-variance to be different between tissues by taking $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{\phi}_h = (\sigma^2_{h1}, \ldots, \sigma^2_{ht})^{\mathrm{T}}$ with respect to the $t = 20$ tissues. We first applied our method to cluster the genes into $g = 4$ groups. The clusters so formed are then compared with the four categories in the GO listings. The adjusted Rand index (Hubert and Arabie, 1985) is adopted to assess the degree of agreement between our partition and the four functional categories. A larger adjusted Rand index indicates a higher level of agreement (Yeung *et al.*, 2003). In our study, the adjusted Rand index was found to be 0.978, which is the best match (the largest index) compared with several model-based and hierarchical clustering algorithms considered in Yeung *et al.* (2003).

We then fit the random-effects model with various number of components *g*. Model selection via BIC indicated that there are seven clusters. The distribution table of the seven clusters compared with the four functional categories in the GO listings is given in Table 2. From Table 2, it can be seen that our clusters 1 and 2 consist mainly of those genes in Categories 2 and 4, respectively. Genes in Category 1 are split into two clusters (4 and 7), while those in Category 3 are separated into three clusters (3, 5 and 6). These subdivisions of functional categories could be relevant to some unknown gene functions in the GO listings.

## 5.3 Clustering of genes on basis of cross-sectional data

This dataset has been used to identify genes that are linked to malignancy of human colorectal carcinoma and is available from the online version of Muro *et al.* (2003). Strictly speaking, this does not represent a microarray experiment as the gene-expression levels are measured based on an adaptor-tagged competitive polymerase chain reaction (PCR) system. We, however, include it here as an example of clustering correlated genes based on measurements

**Table 2.** Distribution of four functional categories over seven clusters obtained (Yeast galactose data)

| Cluster | Category 1 | Category 2 | Category 3 | Category 4 |
|---|---|---|---|---|
| 1 | 0 | 13 | 0 | 0 |
| 2 | 0 | 0 | 0 | 14 |
| 3 | 0 | 2 | 44 | 0 |
| 4 | 38 | 0 | 0 | 0 |
| 5 | 0 | 0 | 17 | 0 |
| 6 | 0 | 0 | 32 | 0 |
| 7 | 45 | 0 | 0 | 0 |

from both tumor and normal tissues. The dataset consists of the expression levels of 1536 genes (known to be associated with colorectal cancers) in 100 tumors and 11 normal tissues ($m = 111$) from which $n = 341$ genes were selected after filtering out poor quality data. Each column of the array data was standardized to have mean zero and unit standard deviation and then each row of the consequent array was standardized. With reference to (1), we take $p = q_e = 2$ corresponding to the above two tissue groups and take $\boldsymbol{X}$ to be an $111 \times 2$ zero-one matrix where the first 100 rows are (1 0) and the next 11 rows are (0 1). The design matrix $\boldsymbol{U}$ is set eqaul to $\boldsymbol{X}$, while $\boldsymbol{V}$ is set to $\boldsymbol{I}_{111}(q_c = m = 111)$. In this study, we allow the $h$-th component-variance to be different between tumor and normal tissues by taking $\boldsymbol{W} = \boldsymbol{X} = \mathbf{X}$ and $\boldsymbol{\phi}_h = (\sigma^2_{h1}, \sigma^2_{h2})^T$ with respect to the two tissue types. Model selection via BIC indicated that there are three clusters. Muro *et al*. (2003) adopted a parametric clustering method within the variational Bayesian framework of Attias (2000) to cluster the genes based on the first three principal components of the data. Their analysis revealed also three clusters. In Figure 2, the first two principal component scores of the clusters obtained by both methods are displayed for comparison. It can be seen that the clusters obtained by both methods are in general well separated in the first two principal components. This is not surprising though for the clusters obtained by Muro *et al*. (2003) as they are formed on the basis of the first three principal components.

As mentioned in Muro *et al*. (2003), their smallest cluster (containing 27 genes) is the most relevant in that it contains genes linked to malignancy. From Figure 2, it can be seen that our method gives a different clustering result for this smallest group (marked as triangles in the figure) to that obtained by Muro *et al*. (2003). They selected a set of 17 representative genes (out of that cluster of 27 genes) to form their tumor-classifier (TCL) genes, which were used to separate the patients into good- and poor-prognosis groups, where the latter group was associated with distant metastases and over-expression of the TCL genes. In our study, we identified a smaller cluster which was a subset of the TCL genes listed in (Muro *et al*., 2003). But as it consisted of only 15 genes, there is no need to reduce it further (as done subjectively in (Muro *et al*., 2003) before forming a classifier to assign the tissue samples into good- and poor-prognosis groups (McLachlan *et al*., 2002). The associations between the tissue groups so formed and the clinical outcomes for the patients such as survival times and the presence of distant metastases can be examined as described in Ben-Tovim Jones *et al*. (2005).

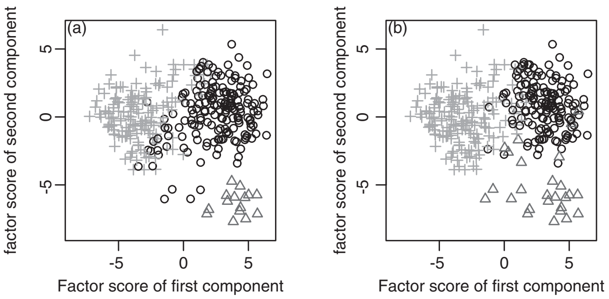

**Fig. 2.** The first two principal scores (colorectal carcinoma data). **(a)** Clusters based on the proposed random-effects model; **(b)** clusters obtained by Muro *et al*. (2003). A colour version of this figure is available as Supplementary material.

## 5.4 Clustering of synthetic time-course data

To illustrate the application of our mixture model in the case where we know the true clustering of the gene profiles, we considered a synthetic time-course dataset. It is based on the yeast cell-cycle data of Spellman *et al*. (1998) as analyzed above. We let $\widehat{\boldsymbol{\Psi}}$ denote the estimate of the vector $\boldsymbol{\Psi}$ of unknown parameters obtained from our analysis in Section 5.1. We generated $n = 600$ observations from our model (1) where, conditional on the cluster-specific random effects, the data were generated from a mixture of $g = 12$ components in proportions $\pi_1, \ldots, \pi_g$. The same fixed-effects structure (11) was imposed on the component-means of the gene profiles as in our original analysis. The values of the parameters in $\boldsymbol{\Psi}$, including the mixing proportions $\pi_i$, were taken to be equal to $\widehat{\boldsymbol{\Psi}}$. We performed 10 simulation trials and on each trial we ran our program EMMIX-WIRE with $g = 12$ components to produce a clustering of the 600 gene profiles into 12 clusters. We computed the adjusted Rand Index for the clustering relative to the true grouping of the 600 (synthetic) genes for each simulation trial, which gave an average value of $\bar{R} = 0.87$. We also ran $k$-means and some agglomerative hierarchical algorithms on each trial for 12 clusters. On each trial, we found that the agglomerative hierarchical clusterings produced by the latter were inferior to $k$-means, regardless of their adopted forms (single, complete and average linkage with the Euclidean or the correlation coefficient metric). A similar conclusion was reached in the much more extensive study of Gibbons and Roth *(*2002*)*. For our simulated trials, the average adjusted Rand index $\bar{R}$ was equal to 0.54 for $k$-means and 0.24 for the agglomerative hierarchical method using complete linkage with the correlation metric.

We also considered the (non-parametric) clustering algorithm CAST proposed by Ben-Dor (1999), using Euclidean distance with the tuning parameter set so as to give the number of clusters as close as possible to 12 as with EMMIX-WIRE. However, it tended to put the majority of the genes into one cluster. For example, for the dataset on the first trial, CAST with the tuning parameter set equal to 0.45, which gave 10 clusters, put 589 of the 600 genes in one cluster with 7 singleton clusters, and 2 other clusters consisting of 2 genes each. We observed a similar result with CAST when applied to the actual yeast cell-cycle data of Spellman *et al*. (1998). The threshold 0.7 was chosen as it gave the closest number of clusters (11) to 12 as with EMMIX-WIRE. We found that CAST put a majority (557) of the 600 genes into 1 cluster with 4 singleton clusters; the other 6 clusters contained 13, 12, 11, 6, 5 and 4 genes.

Finally, we also fitted the standard mixture model with $g = 12$ components with equal covariance matrices to these synthetic datasets. We used the EMMIX program of McLachlan *et al*. (1999) for the fitting of the normal mixture model without structure on the component means and covariance matrices (using 25 random and 25 $k$-means-based starts). It gave an average value of $\bar{R} = 0.32$ for the adjusted Rand index. When we used EMMIX to fit the standard normal mixture model with $g = 12$ components with equal covariance matrices to the actual yeast-cell cycle data of Spellman *et al*. (1998), it gave a clustering that is quite different to that produced by EMMIX-WIRE. The adjusted Rand index of the EMMIX clustering relative to that produced by EMMIX-WIRE is only 0.15.

## 6 DISCUSSION

As an increasing number and a variety of high-throughput datasets become available, cluster analysis is playing an ever increasing role in the analysis of these biological data. The aim of clustering the profile vectors of a very large number of genes is to study the changes in gene expression of entire groups of genes as a means to finding possible functional relationships among them, the identification of transcription factor binding sites and the elucidation of biological pathways. The biological rationale underlying the clustering of the gene profiles is the fact that often many coexpressed genes are also coregulated, which is supported both by an immense body of empirical observations and by detailed mechanistic explanation (Boutros and Okey 2005).

With the analysis of microarray data, there is a need for clustering procedures that can handle data that are both replicated and correlated. We formulate a random-effects model that extends the application of normal mixture models to data arising from a wide variety of experimental designs. Moreover, the model allows for correlations among the gene profile vectors by taking genes within the same cluster to be correlated. We show that this model is able blue provide the program EMMIX-WIRE that enables this model to be fitted very quickly by maximum likelihood via the EM algorithm with the E- and M-steps able to be carried out in closed form.

The proposed method is demonstrated on three representative datasets in the microarray literature blue and also a synthetic dataset. The aim here is not to provide a detailed analysis of these sets, but rather to highlight the potential role and usefulness of our random-effects model for mixture model-based clustering of correlated gene-expression data arising from various biological microarray experiments.

## ACKNOWLEDGEMENT

This work was supported in part by grants from the Australian Research Council and its Centre for Complex Systems, University of Queensland.

*Conflict of Interest*: none declared.

## REFERENCES

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Attias,H. (2000) A variational Bayesian framework for graphical models. In Solla,S.A., Leen,T.K. and Müller,K.R. (eds), *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, pp. 206–212.

Ben-Dor,A., Shamir,R., and Yakhini,Z. (1999) Clustering gene expression patterns. *J. Comput. Biol*., **6**, 281–297.

Ben-Tovim Jones,L., Ng,S.K. and Ambroise,C. (2005) Use of microarray data via model-based classification in the study and prediction of survival from lung cancer. In Shoemaker,J.S. and Lin,S.M. (eds), *Methods of Microarray Data Analysis IV*. Springer, NY, pp. 163–173.

Booth,J.G., Casella,G., Cooke,J.E.K. and Davis,J.M. (2004) Statistical approaches to analysing microarray data representing periodic biological processes: a case study using the yeast cell cycle. Technical report, Department of Biological Statistics and Computational Biology, Cornell University, NY.

Boutros,P.C. and Okey,A.B. (2005) Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. *Brief Bioinform*, **6**, 331–343.

Celeux,G. *et al*. (2005) Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Stat. Model*., **5**, 243–267.

Cheng,J. *et al*. (2004) A knowledge-based clustering algorithm driven by gene ontology. *J. Biopharm. Stat*., **14**, 687–700.

Clare,A. and King,R.D. (2002) How well do we understand the clusters in microarray data? *In Silico Biol*., **2**, 511–522.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B*, **39**, 1–38.

Efron,B. and Tibshirani,R. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, London.

Fraley,C. and Raftery,A.E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis *Comp J*., **41**, 578–588.

Ghosh,D. and Chinnaiyan,A.M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275–286.

Gianola,D. *et al*. (2004) Mixture model for inferring susceptibility to mastitis in diary cattle: a procedure for likelihood-based inference. *Genet. Sel. Evol*., **36**, 3–27.

Gibbons,F.D. and Roth,F.P. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res*., **12**, 1574–1581.

Goldstein,H. (1995) *Multilevel Statistical Models*, (2nd edn). Arnold, London.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif*., **2**, 193–218.

Ideker,T. *et al*. (2001) Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. *Science*, **292**, 929–934.

Klebanov,L. *et al*. (2006) A new type of stochastic dependence revealed in gene expression data. *Stat. Appl. Genetics Mol. Biol*., **5** (Issue 1).

Lee,M.L.T. *et al*. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9838.

Luan,Y. and Li,H. (2003) Clustering of time-course gene expression data using a mixed-effects model with *B*-splines. *Bioinformatics*, **19**, 474–482.

McCulloch,C.E. and Searle,S.R. (2001) *Generalized, Linear, and Mixed Models*. Wiley, NY.

McLachlan,G.J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Stat*., **36**, 318–324.

McLachlan,G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, NY.

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

McLachlan,G.J. *et al*. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

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

McLachlan,G.J. and Khan,N. (2004) On a resampling approach for tests on the number of clusters with mixture model-based clustering of tissue samples. *J. Multivar. Anal*., **90**, 90–105.

McLachlan,G.J. and Peel,D. *Finite Mixture Models*. Wiley, NY.

McLachlan,G.J. *et al*. (1999) The EMMIX software for the fitting of mixtures of normal and *t*-components. *J. Stat. Software*, **4**.

Medvedovic,M. and Sivaganesan,S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.

Muro,S. *et al*. (2003) Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome Biol*., **4**, R21.

Ng,S.K., Krishnan,T. and McLachlan,G.J. (2004) The EM algorithm. In Gentle,J., Hardle,W. and Mori,Y. (eds), *Handbook of Computational Statistics Vol. 1*. Springer-Verlag, NY, pp. 137–168.

Ødegård,J. *et al*. (2005) A Bayesian threshold-normal mixture model for analysis of a continuous mastitis-related trait. *J. Dairy Sci*., **88**, 2652–2659.

Pan,W. (2006) Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, **22**, 795–801.

Pan,W. *et al*. (2002) Model-based cluster analysis of microarray gene-expression data. *Genome Biol*., **3**, research0009.1-0009.8.

Pavlidis,P. *et al*. (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat*., **6**, 461–464.

Spellman,P. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Storey,J.D. *et al*. (2005) Significance analysis of time course microarray experiments. *Proc. Natl Acad. Sci. USA*, **102**, 12837–12842.

Yeung,K.Y. *et al*. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Yeung,K.Y. *et al*. (2003) Clustering gene-expression data with repeated measurements. *Genome Biol*., **4**, R34.