Clustering Replicated Microarray Data via Mixtures of Random Effects Models for Various Covariance Structures

S.K. Ng[1], G.J. McLachlan[1,2], R.W. Bean[1,2], and S.-W. Ng[3]

[1] Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia
Email: {skn, gjm, rbean}@maths.uq.edu.au
[2] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia
[3] Laboratory of Gynecologic Oncology, Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Boston, MA 02115, USA
Email: sng@rics.bwh.harvard.edu

Abstract

A unified approach of mixed-effects model has been recently proposed for clustering correlated genes from different kinds of microarray experiments. With the so-called **EM**-based **MIX**ture analysis **WI**th **R**andom **E**ffects (EMMIX-WIRE) model, both the gene-specific and tissue-specific random effects are taken into account in the (mixture) modelling of microarray data. In this paper, we focus on the applications of the EMMIX-WIRE model to the cluster analysis of microarray data with repeated measurements. In particular, we investigate various forms of covariance structure commonly applicable for replicated microarray data and compare their impact on the final clustering results, using a real data set of microRNA profile and a published yeast galactose data set with known Gene Ontology (GO) listings.

Keywords: EMMIX-WIRE model, Random effects models, Covariance structures, Replicated microarray data.

1 Introduction

The advent of high-throughput technologies has revolutionized molecular biology, and indeed is setting the stage for the rapid evolution of the way disease is diagnosed, classified, and treated. The complexity of tumours makes it likely that a diagnostic test will be based on marker profiles rather than individual markers. However, the identification of relevant subsets of the markers has its challenges, because 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, Kuo, Whitmore & Sklar 2000, Pavlidis, Li & Noble 2003). Replicated measurements of gene expression for a microarray experiment are often correlated and tend to be more alike in characteristics than measurements for the microarray experiments as a whole. At the same time, gene expression levels from the same experiment are correlated (McLachlan, Do & Ambroise 2004). It means that clustering methods which assume independently distributed gene profiles should produce less reliable results than those that exploit or allow for correlation between the gene profiles. Indeed, ignoring the dependence between the gene profiles and the covariance structure of replicated microarray 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 (McLachlan et al. 2004, Ng, McLachlan, Wang, Ben-Tovim & Ng 2006).

Mixed-effects models have been used in the model-based cluster analysis of gene expression data from time-course experiments and experiments with repeated measurements (Luan & Li 2003, Celeux, Martin & Lavergne 2005). However, with these mixed-effects models, only the correlation between replicated measurements for a gene from each microarray experiment is considered (by modelling via gene-specific random effects). Thus, these models require the independence assumption for the genes which, however, will not hold in practice for all pairs of genes (McLachlan et al. 2004, Klebanov, Jordan & Yakovlev 2006) because of the correlation between gene expression levels from the same microarray experiment (tissue-specific effects). Recently, a unified approach of mixed-effects model has been proposed for clustering correlated genes from different kinds of microarray experiments, where both the gene-specific and tissue-specific random effects (Ng et al. 2006) are taken into account in the (mixture) modelling of microarray data. With this so-called **EM**-based **MIX**ture analysis **WI**th **R**andom **E**ffects (EMMIX-WIRE) approach, the unknown model parameters can be obtained by maximum likelihood (ML) via the Expectation-Maximization (EM) algorithm of Dempster *et al.* (1977). see also Ng, Krishnan & McLachlan (2004).

In this paper, we focus on applications of the EMMIX-WIRE procedure to the cluster analysis of microarray data with repeated measurements. In particular, we investigate various forms of covariance structure commonly used for replicated microarray data and compare their impacts on the final clustering results. The rest of the paper is organized as follows: Section 2 introduces the EMMIX-WIRE model for clustering microarray data with repeated measurements and outlines the ML estimation via the EM algorithm. In Section 3, various forms of covariance structure for replicated microarray data are considered and discussed. The impact of various covariance structures on the cluster analysis is studied in Section 4, using a real data set of microRNA profile and a published yeast galactose data set with known Gene Ontology (GO) listings (Ashburner et al. 2000). Section 5 ends the paper with some discussion.

This work was supported by grant from the Australian Research Council.

2 EMMIX-WIRE model for cluster analysis

The EMMIX-WIRE procedure of Ng *et al.* (2006) formulates a linear mixed-effects model (LMM) (McCulloch & Searle 2001) for the mixture components in which covariate information can be incorporated into the clustering process. For t biological samples (not all necessarily independent) with r replicate hybridizations for each, we let $\boldsymbol{y}_j = (\boldsymbol{y}_{1j}^T, \ldots, \boldsymbol{y}_{tj}^T)^T$ contain the expression levels for the jth gene, where

$$\boldsymbol{y}_{lj} = (y_{l1j}, \ldots, y_{lrj})^T \qquad (l = 1, \ldots, t)$$

contains the r technical replicates for the lth biological sample on the jth gene. The superscript T above denotes vector transpose. It is assumed that the (logged) expression levels have been preprocessed with adjustment for array effects. The microarray data can be therefore represented by an $n \times m$ matrix, where $m = t \times r$ is the dimension of the gene-expression profiles. With the EMMIX-WIRE procedure, 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 ith component of the mixture, the vector $\boldsymbol{y}_j$ for the jth gene ($j = 1, \ldots, n$) follows the model

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

where the elements of $\boldsymbol{\beta}_i$ (an t-dimensional vector) are fixed effects (unknown constants) modelling the conditional mean of $\boldsymbol{y}_j$ in the ith component ($i = 1, \ldots, g$). In (1), $\boldsymbol{b}_{ij}$ (an q_b-dimensional vector) and $\boldsymbol{c}_i$ (an q_c-dimensional vector) represent the unobservable gene- and tissue-specific random effects, respectively. These random effects represent the variation due to the heterogeneity of genes and samples (corresponding to $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}^T, \ldots, \boldsymbol{b}_{in}^T)^T$ and $\boldsymbol{c}_i$, respectively). The random effects $\boldsymbol{b}_i$ and $\boldsymbol{c}_i$, and the measurement error vector $(\boldsymbol{\epsilon}_{i1}^T, \ldots, \boldsymbol{\epsilon}_{in}^T)^T$ are assumed to be mutually independent, where $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are known design matrices of the corresponding fixed or random effects, respectively.

With the LMM, the distributions of $\boldsymbol{b}_{ij}$ and $\boldsymbol{c}_i$ are taken, respectively, to be multivariate normal $N_{q_b}(\boldsymbol{0}, \theta_{bi}\boldsymbol{I}_{q_b})$ and $N_{q_c}(\boldsymbol{0}, \theta_{ci}\boldsymbol{I}_{q_c})$, where $\boldsymbol{I}_{q_b}$ and $\boldsymbol{I}_{q_c}$ are identity matrices with dimensions being specified by the subscripts. The measurement error vector $\boldsymbol{\epsilon}_{ij}$ is also taken to be multivariate normal $N_m(\boldsymbol{0}, \boldsymbol{A}_i)$, where $\boldsymbol{A}_i = \mathrm{diag}(\boldsymbol{W}\boldsymbol{\phi}_i)$ is a diagonal matrix constructed from the vector $(\boldsymbol{W}\boldsymbol{\phi}_i)$ with $\boldsymbol{\phi}_i = (\sigma_{i1}^2, \ldots, \sigma_{iq_e}^2)^T$ and $\boldsymbol{W}$ a known $m \times q_e$ zero-one design matrix. That is, we allow the ith component-variance to be different among the m hybridizations.

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

3 Covariance structures for replicated experiments

Let $\boldsymbol{Y}^i$ denote a random vector of size $n_i m$ consisting of all the observations $\boldsymbol{y}_j$ that arise from the ith component, where n_i is the number of genes belonging to the ith component. It is assumed that all $\boldsymbol{y}_j$ in the ith component are independent given $\boldsymbol{c}_i$. The conditional distribution of $\boldsymbol{Y}^i \mid \boldsymbol{c}_i$ is then given by $N_{n_i m}(\boldsymbol{\Lambda}_i\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Lambda}_i = (\boldsymbol{1}_{n_i} \otimes \boldsymbol{X})$. Here, $\boldsymbol{1}_{n_i}$ is an n_i-dimensional vector of ones, the sign $\otimes$ is the Kronecker product of two matrices, and

$$\boldsymbol{\Sigma}_i = \boldsymbol{I}_{n_i} \otimes (\boldsymbol{A}_i + \theta_{bi}\boldsymbol{U}\boldsymbol{U}^T).$$

Hence, the unconditional distribution of $\boldsymbol{Y}^i$ is given by $N_{n_i m}(\boldsymbol{\Lambda}_i\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i + \boldsymbol{J}_{n_i} \otimes \boldsymbol{D}_i)$, where $\boldsymbol{J}_{n_i}$ is an $n_i \times n_i$ matrix of ones and

$$\boldsymbol{D}_i = \theta_{ci}\boldsymbol{V}\boldsymbol{V}^T. \qquad (2)$$

The presence of the term (2) in the covariance matrix of $\boldsymbol{Y}^i$ induces the correlation between genes that belong to the same cluster.

For the specification of gene-specific random effects $\boldsymbol{b}_{ij}$, we consider two typical models applicable for replicated microarray data. The first model takes $\boldsymbol{U} = \boldsymbol{X}$ and $q_b = t$ such that $\boldsymbol{b}_{ij} = (b_{i1j}, \ldots, b_{itj})^T$. That is, it is assumed that a gene-specific random effect, b_{ilj}, is shared among the repeated measurements of expression on the jth gene in the lth biological sample ($l = 1, \ldots, t$). The replicated measurements are therefore correlated. The second model simplifies the first one by taking $\boldsymbol{U} = \boldsymbol{1}_m$ and $q_b = 1$. That is, it is assumed that a gene-specific random effect, b_{ij}, is shared among the measurements on the jth gene from all the $m = t \times r$ hybridizations.

For the specification of tissue-specific random effects $\boldsymbol{c}_i$, we consider three typical models applicable for replicated microarray data. The first model takes $\boldsymbol{V} = \boldsymbol{I}_m$ and $q_c = m = t \times r$ such that $\boldsymbol{c}_i = (c_{i11}, \ldots, c_{ir1}, \ldots, c_{il1}, \ldots, c_{irt})^T$. That is, it is assumed that a tissue-specific random effect, c_{ikl}, is shared among gene expressions from the kth replicate of the lth biological sample ($k = 1, \ldots, r$; $l = 1, \ldots, t$). It means that genes within the same cluster are correlated. In some microarray experiments, the t biological samples, however, are not all independent. For example, they could correspond to samples from p patients with $t_1 + t_2 + \ldots + t_p = t$. The value t_s corresponds to the number of biological samples from the sth patient ($s = 1, \ldots, p$). For example, the t_s biological samples for the sth patient might correspond to samples taken at t_s different time points or in t_s different conditions. A second model can be adopted to incorporate such a data hierarchy by taking

$$\boldsymbol{V} = \boldsymbol{V}^* = \begin{pmatrix} \boldsymbol{1}_{t_1 r} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1}_{t_2 r} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{1}_{t_p r} \end{pmatrix},$$

and $q_c = p$. It means that a patient-specific random effect, c_{is}, is shared among gene expression levels for the technical and biological replicates for the sth patient ($s = 1, \ldots, p$). It thus induces a correlation between the expression levels of different genes on the same patient provided the genes belong to the same cluster. The third model simplifies the above two models by taking $\boldsymbol{V} = \boldsymbol{0}$. That is, it is assumed that there are no tissue-specific random effects and genes are not correlated (an independence model).

By considering the combinations of the above random-effects models, we have six forms of covariance structures:

Model 1: Taking $\boldsymbol{U} = \boldsymbol{X}$, $q_b = t$, $\boldsymbol{V} = \boldsymbol{I}_m$, $q_c = m$, $\boldsymbol{W} = \boldsymbol{X}$, and $q_e = t$, the covariance matrix for the unconditional distribution of $\boldsymbol{Y}^i$ is given by

$$\boldsymbol{I}_{n_i} \otimes (\mathrm{diag}(\boldsymbol{X}\boldsymbol{\phi}_i) + \theta_{bi}\boldsymbol{X}\boldsymbol{X}^T) + \boldsymbol{J}_{n_i} \otimes \theta_{ci}\boldsymbol{I}_m\boldsymbol{I}_m^T,$$

where $\boldsymbol{\phi}_i = (\sigma_{i1}^2, \ldots, \sigma_{it}^2)^T$.

Model 2: Taking $\boldsymbol{U} = \boldsymbol{X}$, $q_b = t$, $\boldsymbol{V} = \boldsymbol{V}^*$, $q_c = p$, $\boldsymbol{W} = \boldsymbol{X}$, and $q_e = t$, the covariance matrix for $\boldsymbol{Y}^i$ is given by

$$\boldsymbol{I}_{n_i} \otimes (\mathrm{diag}(\boldsymbol{X}\boldsymbol{\phi}_i) + \theta_{bi}\boldsymbol{X}\boldsymbol{X}^T) + \boldsymbol{J}_{n_i} \otimes \theta_{ci}\boldsymbol{V}^*\boldsymbol{V}^{*T},$$

where $\boldsymbol{\phi}_i = (\sigma_{i1}^2, \ldots, \sigma_{it}^2)^T$.

Model 3: Taking $\boldsymbol{U} = \boldsymbol{X}$, $q_b = t$, $\boldsymbol{V} = \boldsymbol{0}$, $\boldsymbol{W} = \boldsymbol{X}$, and $q_e = t$, the covariance matrix for $\boldsymbol{Y}^i$ is given by

$$\boldsymbol{I}_{n_i} \otimes (\mathrm{diag}(\boldsymbol{X}\boldsymbol{\phi}_i) + \theta_{bi}\boldsymbol{X}\boldsymbol{X}^T),$$

where $\boldsymbol{\phi}_i = (\sigma_{i1}^2, \ldots, \sigma_{it}^2)^T$.

Model 4: Taking $\boldsymbol{U} = \boldsymbol{1}_m$ and $q_b = 1$, $\boldsymbol{V} = \boldsymbol{I}_m$, $q_c = m$, $\boldsymbol{W} = \boldsymbol{1}_m$, and $q_e = 1$, the covariance matrix for $\boldsymbol{Y}^i$ is given by

$$\boldsymbol{I}_{n_i} \otimes (\mathrm{diag}(\boldsymbol{1}_m\phi_i) + \theta_{bi}\boldsymbol{1}_m\boldsymbol{1}_m^T) + \boldsymbol{J}_{n_i} \otimes \theta_{ci}\boldsymbol{I}_m\boldsymbol{I}_m^T,$$

where $\phi_i = \sigma_i^2$.

Model 5: Taking $\boldsymbol{U} = \boldsymbol{1}_m$ and $q_b = 1$, $\boldsymbol{V} = \boldsymbol{V}^*$, $q_c = p$, $\boldsymbol{W} = \boldsymbol{1}_m$, and $q_e = 1$, the covariance matrix for $\boldsymbol{Y}^i$ is given by

$$\boldsymbol{I}_{n_i} \otimes (\mathrm{diag}(\boldsymbol{1}_m\phi_i) + \theta_{bi}\boldsymbol{1}_m\boldsymbol{1}_m^T) + \boldsymbol{J}_{n_i} \otimes \theta_{ci}\boldsymbol{V}^*\boldsymbol{V}^{*T},$$

where $\phi_i = \sigma_i^2$.

Model 6: Taking $\boldsymbol{U} = \boldsymbol{1}_m$, $q_b = 1$, $\boldsymbol{V} = \boldsymbol{0}$, $\boldsymbol{W} = \boldsymbol{1}_m$, and $q_e = 1$, the covariance matrix for $\boldsymbol{Y}^i$ is given by

$$\boldsymbol{I}_{n_i} \otimes (\mathrm{diag}(\boldsymbol{1}_m\phi_i) + \theta_{bi}\boldsymbol{1}_m\boldsymbol{1}_m^T),$$

where $\phi_i = \sigma_i^2$.

To examine the (biological) meaning of Equation (1) for the various models above, we consider Model 1. Under this model, it is assumed that the expression level of the jth gene, conditional of its membership of the ith component of the mixture (ith cluster), is given for the kth replicate in the lth experiment by

$$y_{jkl} = \beta_{il} + b_{ilj} + c_{ikl} + \epsilon_{ijkl}$$

($i = 1, \ldots, g$; $j = 1, \ldots, n$; $k = 1, \ldots, r$; $l = 1, \ldots, t$). That is, the expression level y_{jkl} is equal to the mean expression level at the lth experiment for the ith component (β_{il}) plus a gene-specific random effect b_{ilj}, a tissue-specific random effect c_{ikl}, and an experimental random error ϵ_{ijkl}. The vector of dimension $q_b = t$, $(b_{i1j}, \ldots, b_{itj})^T$, represents the variation between the gene expression profiles and their component-means for the t microarray experiments. The vector of dimension $q_c = m$, $(c_{i11}, \ldots, c_{ir1}, \ldots, c_{i1t}, \ldots, c_{irt})^T$, represents the variation between expression signature and the component-mean signature for the $m = t \times r$ hybridizations.

It can be seen from the covariance matrix for $\boldsymbol{Y}^i$ that Models 3 and 6 are independence models, where there are no tissue-specific random effects being assumed ($\boldsymbol{V} = \boldsymbol{0}$). It means that expression levels for the same microarray experiment are independent.

4 Comparative studies

The impact of various covariance structures on the cluster analysis is compared using a real data set of microRNA profile and a published yeast galactose data set with known GO listings.

MicroRNAs are a family of small (~22 nucleotides) noncoding RNA molecules that are evolutionary conserved and are expressed in a tissue-specific and developmental stage-specific manner (Bartel 2004). They are important regulators of various aspects of developmental control in both plants and animals through sequence-specific interactions with target mRNAs. Recent studies have shown that microRNA expression profiles are more accurate than global mRNA profiles in classifying the histologic origins and differentiation of human tumours (Lu et al. 2005) and highlighted the potential of microRNA profiling in cancer diagnosis and classification (He et al. 2005). The data set consists of three ($r = 3$) replicate hybridizations for each microRNA microarray experiment of $t = 12$ samples. However, there is a large amount of missing data. We therefore work with a subset of $n = 160$ microRNAs that have about 13% of the data missing. All the missing expressions were imputed using the support vector regression (SVR) imputation and orthogonal coding scheme (Wang, Li, Jiang & Feng 2006). We are interested primarily in which microRNAs 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 the Bayesian information criterion (BIC) of Schwarz (1978). This criterion, which is based on a penalized form of the log likelihood, has growing support in the literature for selecting the value of g in the context of mixture model-based clustering of microarray data (Luan & Li 2003, Yeung, Fraley, Murua, Raftery & Ruzzo 2001); see also the discussion in Ng *et al.* (2006).

The covariance structures in Models 1, 3, 4, and 6 presented in Section 3 are considered now. Models 2 and 5 were not considered as there was no information available on the experiments to suggest that they would be applicable. As an illustration for Model 1, we take $m = t \times r = 36$ and $\boldsymbol{X} = \boldsymbol{1}_3 \otimes \boldsymbol{I}_{12}$ (a 36×12 matrix). The design matrices $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{W}$ are taken to be equal to $\boldsymbol{X}$, $\boldsymbol{I}_{36}$, and $\boldsymbol{X}$, respectively. We fit this model for various values of the number of components g. Model selection via BIC indicated that there are five clusters.

Based on the setting of $g = 5$, we then fit the mixed-effects models with various covariance structures. The clusters so formed are then compared to that obtained from Model 1 above. The adjusted Rand index (Hubert & Arabie 1985) is adopted to assess the degree of agreement between two clustering partitions. A larger adjusted Rand index indicates a higher level of agreement. Identical clustering partitions will have the adjusted Rand index of one. In Table 1, the adjusted Rand indices for various covariance structures considered are presented. It can be seen that various covariance structures did result in different clustering of microRNAs.

To illustrate further the relative impact of the

Table 1: Adjusted Rand indices with reference to the clustering obtained from Model 1 (MicroRNA data)

Covariance structure	Adjusted Rand index
Model 1	1.0
Model 3	0.723
Model 4	0.298
Model 6	0.298

Table 2: Adjusted rand indices with reference to the known GO listings (Yeast galactose data)

Covariance structure	Adjusted rand index
Model 1	0.978
Model 3	0.811
Model 4	0.906
Model 6	0.910

adopted covariance structure on the cluster analysis, we work on a published yeast data set with known GO listings (Ideker et al. 2001, Yeung, Medvedovic & Bumgarner 2003). With this yeast galactose data, there are four ($r = 4$) replicate hybridizations for each cDNA array experiment. There are $n = 205$ genes and $t = 20$ microarray experiments. The expression patterns of these 205 genes reflect four functional categories in the GO listings (Yeung et al. 2003). We first applied Model 1 given in Section 3 to cluster the genes into $g = 4$ groups. The clusters so formed are then compared to the four categories in the GO listings. 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). The adjusted Rand indices for mixed-effects models with various covariance structures are given in Table 2. Again, it can be seen that different clustering results are obtained from the various covariance structures considered.

5 Discussion

We have investigated various covariance structures in EMMIX-WIRE model applicable for clustering replicated microarray data. The specification of covariance structures needs careful consideration. The choice should be justified by the data hierarchy so formed due to the design of microarray experiments. With repeated measures data, replicated measurements of size r from t microarray experiments on each gene are obtained. It is therefore anticipated that random effects are shared among expression levels to represent the variation due to the heterogeneity of genes and samples (corresponding to $\boldsymbol{b}_i$ and $\boldsymbol{c}_i$, respectively), as discussed in Section 3. It is interesting to note that combinations of random effects may be considered in mixed-effects modelling. For example, an alternative model for the specification of gene-specific random effects $\boldsymbol{b}_{ij}$ may be adopted by combining the two models $\boldsymbol{U} = \boldsymbol{X}$ and $\boldsymbol{U} = \boldsymbol{1}_m$ together (that is, a random effect accounting for correlation among replicated measurements plus another accounting for correlation among all hybridizations). However, it was demonstrated in Celeux *et al.* (2005) that this model provided quite similar results to that of the first model with $\boldsymbol{U} = \boldsymbol{X}$. This result indicates that combinations of random effects are usually not required.

The impact of various covariance structures on the clustering results are compared in Section 4. It can be seen that Model 1 outperforms others for the cluster analysis of the yeast galactose data. With Model 1, it is assumed that a gene-specific random effect, b_{ilj}, is shared among the repeated measurements on the jth gene from the lth microarray experiment ($l = 1, \ldots, t$). A tissue-specific random effect, c_{ikl}, is also assumed to be shared among gene expressions from the kth replicate for the lth experiment ($k = 1, \ldots, r; l = 1, \ldots, t$). It means that replicated measurements are correlated and genes within the same cluster are also correlated. This correlation structure is justified by the data hierarchy so formed in typical replicated microarray experiments. On the other hand, the simplified model for the specification of $\boldsymbol{b}_{ij}$ with $\boldsymbol{U} = \boldsymbol{1}_m$ can be regarded as unrealistic in many situations of replicated microarray experiments (Celeux et al. 2005), as indicated in Table 1 for the microRNA data.

References

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

Bartel, D.P. (2004), 'MicroRNAs: genomics, biogenesis, mechanism, and function', *Cell* **116**, 281–297.

Celeux, G., Martin, O. & Lavergne, C. (2005), 'Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments', *Statistical Modelling* **5**, 243–267.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *J. Roy. Stat. Soc. Ser. B* **39**, 1–38.

He, L., Thomson, J.M., Hemann, M.T., et al. (2005), 'A microRNA polycistron as a potential human oncogene', *Nature* **435**, 828–833.

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

Ideker, T., Thorsson, V., Ranish, J.A., et al. (2001), 'Integrated genomic and proteomic analyses of a systemically perturbed metabolic network', *Science* **292**, 929–934.

Klebanov, L., Jordan, C. & Yakovlev, A. (2006), 'A new type of stochastic dependence revealed in gene expression data', *Stat. Appl. Genetics Mol. Biol.* **5**, No. 1, Article 7.

Lee, M.L.T., Kuo, F.C., Whitmore, G.A. & Sklar, J. (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.

Lu, Y., Getz, G., Miska, E.A., et al. (2005), 'MicroRNA expression profiles classify human cancers', *Nature* **435**, 834–838.

Luan, Y. & 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. & Searle, S.R. (2001), *Generalized, Linear, and Mixed Models*, Wiley.

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

Ng, S.K., Krishnan, T. & McLachlan, G.J. (2004), The EM algorithm, *in* J. Gentle, W. Hardle & Y. Mori, eds, 'Handbook of Computational Statistics Vol. 1', Springer-Verlag, pp. 137–168.

Ng, S.K., McLachlan, G.J., Wang, K., Ben-Tovim, L. & Ng, S.-W. (2006), 'A mixture model with random-effects components for clustering correlated gene-expression profiles', *Bioinformatics* **22**, 1745–1752.

Pavlidis, P., Li, Q. & Noble, W.S. (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.

Wang, X., Li, A., Jiang, Z.H. & Feng, H.Q. (2006), 'Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme', *BMC Bioinformatics* **7**, Article 32.

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

Yeung, K.Y., Medvedovic, M. & Bumgarner, R.E. (2003), 'Clustering gene-expression data with repeated measurements', *Genome Biol.* **4**, Article R34.