BIOINFORMATICS

Vol. 18 no. 3 2002
Pages 413–422

A mixture model-based approach to the clustering of microarray expression data

G. J. McLachlan, R. W. Bean and D. Peel

Department of Mathematics, University of Queensland, Brisbane, Queensland 4072, Australia

Received on August 30, 2001; revised on October 26, 2001; accepted on November 2, 2001

ABSTRACT

Motivation: This paper introduces the software EMMIX-GENE that has been developed for the specific purpose of a model-based approach to the clustering of microarray expression data, in particular, of tissue samples on a very large number of genes. The latter is a nonstandard problem in parametric cluster analysis because the dimension of the feature space (the number of genes) is typically much greater than the number of tissues. A feasible approach is provided by first selecting a subset of the genes relevant for the clustering of the tissue samples by fitting mixtures of t distributions to rank the genes in order of increasing size of the likelihood ratio statistic for the test of one versus two components in the mixture model. The imposition of a threshold on the likelihood ratio statistic used in conjunction with a threshold on the size of a cluster allows the selection of a relevant set of genes. However, even this reduced set of genes will usually be too large for a normal mixture model to be fitted directly to the tissues, and so the use of mixtures of factor analyzers is exploited to reduce effectively the dimension of the feature space of genes.

Results: The usefulness of the EMMIX-GENE approach for the clustering of tissue samples is demonstrated on two well-known data sets on colon and leukaemia tissues. For both data sets, relevant subsets of the genes are able to be selected that reveal interesting clusterings of the tissues that are either consistent with the external classification of the tissues or with background and biological knowledge of these sets.

Availability: EMMIX-GENE is available at http://www.maths.uq.edu.au/~gjm/emmix-gene/

Contact: gjm@maths.uq.edu.au

1 INTRODUCTION

The analysis of gene expression microarray data using clustering techniques has an important role to play in the discovery, validation, and understanding of various classes and subclasses of cancer; see, for example, Eisen *et al.* (1998), Ben-Dor *et al.* (1999, 2000), Alon *et al.* (1999), Golub *et al.* (1999), Hastie *et al.* (2000), Moler *et al.* (2000), Nguyen and Rocke (2001), and Xing and Karp (2001), among others. The clustering algorithm we present here, called EMMIX-GENE, can be applied to the problem of clustering tissue samples on the basis of genes and to the problem of clustering genes on the basis of tissues. For the clustering of genes, the EMMIX-GENE software makes use of existing options from the EMMIX program of McLachlan *et al.* (1999). The tissue space and the gene space are generally of quite different dimensionality (10–10^2 tissues versus 10^3–10^4 genes). The clustering of the genes on the basis of the tissues is therefore a standard cluster analysis problem that can be effected by using existing software to fit normal mixture models. But unless the genes are assumed to be uncorrelated within a cluster, the clustering of the tissue samples on the basis of all the genes is nonstandard since the dimension of each tissue sample (the number of genes) is so much greater than the number of tissues. This dimensionality problem is handled with the EMMIX-GENE approach by fitting mixtures of factor analyzers, which allow for nonzero component-correlations between the genes. Given the very large number of genes in a typical tissue sample, EMMIX-GENE initially considers a reduction in the number of genes to be used in the clustering process.

The EMMIX-GENE approach is to be illustrated in the clustering of two well-known data sets in the microarray literature, the colon data analyzed initially in Alon *et al.* (2000), and the leukaemia data first analyzed in Golub *et al.* (1999).

2 NORMAL MIXTURE MODELS

Before we proceed to present the EMMIX-GENE approach, we shall briefly summarize the normal mixture model and the extensions to mixtures of t distributions and to mixtures of factor analyzers. Finite mixtures of distributions have provided a sound mathematical-based approach to the statistical modelling of a wide variety of random phenomena; see, for example, McLachlan and Peel (2000a). For multivariate data of a continuous nature,

attention has focused on the use of multivariate normal components because of their computational convenience. We let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ denote n p-dimensional observations. With a normal mixture model-based approach to clustering of these data, it is assumed that each observation $\boldsymbol{x}_j$ is from a mixture of an initially specified number g of multivariate normal densities in some unknown proportions $\pi_1, \ldots, \pi_g$. That is, $\boldsymbol{x}_j$ is taken to be a realization of a random vector $\boldsymbol{X}$ having the mixture probability density function (p.d.f.) $f(\boldsymbol{x}; \boldsymbol{\Psi})$ defined by,

$$f(\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where $\phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p-variate normal density probability function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i (i = 1, \ldots, g)$. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions π_i, the elements of the component means $\boldsymbol{\mu}_i$, and the distinct elements of the component–covariance matrices $\boldsymbol{\Sigma}_i (i = 1, \ldots, g)$.

Under the assumption that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent observations, the log likelihood function for the parameter vector $\boldsymbol{\Psi}$ can be formed by summing over the log mixture density at each point $\boldsymbol{x}_j$ to give

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log f(\boldsymbol{x}_j; \boldsymbol{\Psi}). \tag{2}$$

The maximum likelihood estimate of $\boldsymbol{\Psi}$ is obtained as an appropriate root of the likelihood equation

$$\partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = \boldsymbol{0}. \tag{3}$$

Solutions of (3) corresponding to local maxima can be found iteratively by application of the Expectation–Maximization (EM) algorithm of Dempster *et al.* (1977); see also McLachlan and Krishnan (1997). The EM algorithm is applied in the framework where each observation $\boldsymbol{x}_j$ is conceptualized to have arisen from one of the components and the indicator variable denoting its component of origin is taken to be missing. The so-called complete-data log likelihood is formed on the basis of these indicator variables in addition to the observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. On the E-step, the complete-data log likelihood is averaged over the conditional distribution of the indicator variables given the observed data, using the current estimate of the parameter vector. Since the complete-data log likelihood is linear in these indicator variables, the E-step of the EM algorithm simply involves replacing them by the current values of their conditional expectations, which are the so-called posterior probabilities of component membership. The posterior probability that the jth data point belongs to the ith component of the mixture is written here as $\tau_i(\boldsymbol{x}_j; \boldsymbol{\Psi})$ and is given by

$$\tau_i(\boldsymbol{x}_j; \boldsymbol{\Psi}) = \pi_i \phi(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) / f(\boldsymbol{x}_j; \boldsymbol{\Psi})$$

for $i = 1, \ldots, g$ and $j = 1, \ldots, n$. On the M-step, the estimates of the component mixing proportions, means, and covariance matrices are updated by using the current values for the posterior probabilities in place of the indicator variables in the usual closed-form expressions for the sample proportions, means, and covariance matrices. The E- and M-steps are alternated repeatedly until convergence of the EM sequence of iterates. The EM algorithm has reliable global convergence in that regardless of the starting point, the likelihood (2) is increased after each EM iteration and that convergence is to a local maximum, assuming that the process is not attracted to a spike in the likelihood function.

Once the mixture model has been fitted, a probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data, $\tau(\boldsymbol{x}_j; \hat{\boldsymbol{\Psi}})$, where $\hat{\boldsymbol{\Psi}}$ denotes the maximum likelihood estimate of $\boldsymbol{\Psi}$. An outright assignment of the data into g clusters is achieved by assigning each data point to the component to which it has the highest estimated posterior probability of belonging. The likelihood ratio statistic λ can be used to test for the smallest number of components in the mixture model compatible with the data. However, the situation is not straightforward since regularity conditions do not hold for the asymptotic null distribution of $-2 \log \lambda$ to be chi-squared; nor do they hold for the justification of the Bayesian Information Criterion (BIC), although it still appears to provide a useful informal guide in practice (McLachlan and Peel, 2000a, Chapter 6). A formal test can be carried out using a resampling approach as proposed in McLachlan (1987).

2.1 Mixtures of t distributions

The use of t component distributions is employed in the gene-selection stage of the EMMIX-GENE program in order to provide some protection against atypical observations, which are prevalent in microarray data. With the t mixture model-based approach, the normal distribution for the ith component in the mixture is embedded in a wider class of elliptically symmetric distributions with an additional parameter called the degrees of freedom ν_i. As ν_i tends to infinity, the t distribution approaches the normal distribution. Hence this parameter ν_i may be viewed as a robustness tuning parameter. It can be fixed in advance or it can be inferred from the data for each component thereby providing an *adaptive* robust procedure.

The t density with location parameter $\boldsymbol{\mu}_i$, positive definite inner product matrix $\boldsymbol{\Sigma}_i$, and ν_i degrees of freedom is given by

$$\frac{\Gamma(\frac{\nu_i + p}{2}) |\boldsymbol{\Sigma}_i|^{-1/2}}{(\pi \nu_i)^{\frac{1}{2}p} \Gamma(\frac{\nu_i}{2}) \{1 + \delta(\boldsymbol{x}, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)/\nu_i\}^{\frac{1}{2}(\nu_i + p)}},$$

where

$$\delta(\boldsymbol{x}, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i) = (\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)$$

denotes the Mahalanobis squared distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_i$. If $\nu_i > 1$, $\boldsymbol{\mu}_i$ is the mean of $\boldsymbol{X}$, and if $\nu_i > 2$, $\nu_i(\nu_i - 2)^{-1}\boldsymbol{\Sigma}_i$ is its covariance matrix.

McLachlan and Peel (2000a, Chapter 7) have provided a detailed account how the EM algorithm and a multicycle Expectation–Conditional Maximization (ECM) variant can be used to undertake maximum likelihood estimation of a mixture of t distributions with unspecified degrees of freedom ν_i. If ν_i is fixed in advance for each component, then the M-step exists in closed form with the component means and covariance matrices updated effectively using weighted least squares.

3 MIXTURES OF FACTOR ANALYZERS

3.1 Single-component factor model

Factor analysis is commonly used for explaining correlations between variables in multivariate observations. It can be used also for dimensionality reduction. In a typical factor analysis model, each observation $\boldsymbol{X}_j$ is modelled as

$$\boldsymbol{X}_j = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{U}_j + \boldsymbol{e}_j \quad (j = 1, \ldots, n), \tag{4}$$

where $\boldsymbol{U}_j$ is a q-dimensional ($q < p$) vector of latent or unobservable variables called factors and $\boldsymbol{B}$ is a $p \times q$ matrix of factor loadings (parameters). The $\boldsymbol{U}_j$ are assumed to be independent and identically (i.i.d.) as $N(\boldsymbol{0}, \boldsymbol{I}_q)$, independently of the errors $\boldsymbol{e}_j$, which are assumed to be i.i.d. as $N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a diagonal matrix,

$$\boldsymbol{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2),$$

and where $\boldsymbol{I}_q$ denotes the $q \times q$ identity matrix. Thus, conditional on the u_j, the $\boldsymbol{X}_j$ are independently distributed as $N(\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{u}_j, \boldsymbol{D})$. Unconditionally, the $\boldsymbol{X}_j$ are i.i.d. according to a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D}. \tag{5}$$

If q is chosen sufficiently smaller than p, the representation (5) imposes some constraints on the component–covariance matrix $\boldsymbol{\Sigma}$ and thus reduces the number of free parameters to be estimated. Note that in the case of $q > 1$, there is an infinity of choices for $\boldsymbol{B}$, since (5) is still satisfied if $\boldsymbol{B}$ is replaced by $\boldsymbol{B}\boldsymbol{C}$, where $\boldsymbol{C}$ is any orthogonal matrix of order q. One (arbitrary) way of uniquely specifying $\boldsymbol{B}$ is to choose the orthogonal matrix $\boldsymbol{C}$ so that $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{B}$ is diagonal (with its diagonal elements arranged in decreasing order). Assuming that the eigenvalues of $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ are positive and distinct, the condition that $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{B}$ is diagonal as above imposes $\frac{1}{2}q(q-1)$ constraints on the parameters. Hence then the number of free parameters is $pq + p - \frac{1}{2}q(q-1)$.

With the factor analysis model (4), we avoid having to compute the inverses of iterates of the estimated $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ that may be singluar for large p relative to n. This is because the inversion of the current value of the $p \times p$ matrix $(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D})$ on each iteration can be undertaken using the result that

$$(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D})^{-1} = \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{B} \times (\boldsymbol{I}_q + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}, \tag{6}$$

where the right-hand side of (6) involves only the inverses of $q \times q$ matrices, since $\boldsymbol{D}$ is a diagonal matrix. The determinant of $(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D})$ can then be calculated as

$$|\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D}| = |\boldsymbol{D}|/|\boldsymbol{I}_q - \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D})^{-1}\boldsymbol{B}|.$$

Unlike the principal components model, the factor analysis model (4) enjoys a powerful invariance property: changes in the scales of the feature variables in $\boldsymbol{x}_j$, appear only as scale changes in the appropriate row of the matrix $\boldsymbol{B}$ of factor loadings.

3.2 Mixtures of factor models

As the single-factor analysis model (4) provides only a global linear model for the representation of the data in a lower-dimensional subspace, the scope of its application is limited. A global nonlinear approach can be obtained by postulating a finite mixture of linear submodels for the distribution of the full observation vector $\boldsymbol{X}_j$ given some (unobservable) factors, as advocated in McLachlan and Peel (2000a,b). This model was originally proposed by Ghahramani and Hinton (1997) for the purposes of visualizing high dimensional data in a lower dimensional space to explore for group structure.

The mixture of factor analyzers model is given by (1), where now the ith component–covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{B}_i^{\mathrm{T}} + \boldsymbol{D}_i \quad (i = 1, \ldots, g), \tag{7}$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix ($i = 1, \ldots, g$). The parameter vector $\boldsymbol{\Psi}$ now consists of the elements of the $\boldsymbol{\mu}_i$, the $\boldsymbol{B}_i$, and the $\boldsymbol{D}_i$, along with the mixing proportions $\pi_i (i = 1, \ldots, g)$. McLachlan and Peel (2000a, Chapter 8) have described how a variant of the EM algorithm, the Alternating Expectation–Conditional Maximization (AECM) algorithm of Meng and van Dyk (1997), can be used to fit the mixture of factor analyzers by maximum likelihood.

At the final values of the iterates for the parameters, the maximum likelihood estimate of the diagonal matrix $\boldsymbol{D}_i$ satisfies

$$\hat{\boldsymbol{D}}_i = \mathrm{diag}(\hat{\boldsymbol{V}}_i - \hat{\boldsymbol{B}}_i\hat{\boldsymbol{B}}_i^{\mathrm{T}}), \tag{8}$$

where $\hat{\boldsymbol{V}}_i$ is the ith component sample covariance matrix with the observations weighted by the final values of the ith component posterior probabilities. It can be seen from (8) that some of the estimates of the elements of the diagonal matrix $\boldsymbol{D}_i$ will be close to zero if effectively not more than q observations are unequivocally assigned to the ith component of the mixture on the basis of the fitted posterior probabilities of component membership. This will lead to spikes or near singularities in the likelihood. One way to avoid this is to impose the condition of a common value $\boldsymbol{D}$ for the $\boldsymbol{D}_i$,

$$\boldsymbol{D}_i = \boldsymbol{D} \quad (i = 1, \ldots, g). \tag{9}$$

4 DIMENSION REDUCTION

In the standard setting of a model-based cluster analysis, the n observations to be clustered are taken to be independent realizations where the sample size n is much larger than the dimension p of each observation,

$$n \gg p. \tag{10}$$

It is also assumed that the sizes of the clusters to be produced are sufficiently large relative to p to avoid any singular estimates of the within-cluster covariance matrices.

We now consider the cluster analysis of microarray data collected on N genes from M experiments, which can be represented in the form of a $N \times M$ data matrix $\boldsymbol{A}$ whose ith row contains the expression levels for the ith gene in the M tissue samples. Typically, N is typically larger than M. Thus for the problem of clustering N genes on the basis of the M tissues, we have $n = N$ and $p = M$, and so condition (10) for a standard cluster analysis will be satisfied usually. The condition of independent data will not hold given that not all the genes in a given tissue sample are independently distributed. But in practice we can proceed with the standard clustering methodology, ignoring any correlations between genes in the same tissue sample.

We now consider the problem of clustering the M tissues on the basis of the N genes. For this problem, we have $n = M$ and $p = N$, and so the sample size n will be typically small relative to the dimension p, thus causing estimation problems with the normal mixture model. This is because the g-component normal mixture model (1) with unrestricted component–covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component–covariance matrix $\boldsymbol{\Sigma}_i (i = 1, \ldots, g)$. It therefore cannot be fitted directly to the tissues on the basis of all the $p = N$ genes. The EMMIX-GENE program handles this high-dimensional problem by using mixtures of factor analyzers, where $\boldsymbol{\Sigma}_i$ is specified by (7) and (9). A reduction in the number of parameters is achieved by taking the number of factors q to be appropriately small. Although the model under (9) can be fitted provided q is less than the sample size n, q needs to be sufficiently small to ensure that the estimates of the component–covariance matrices are not highly variable. Hence q may not be able always to be taken sufficiently large to model adequately the full correlation structure of the genes in the lower q-dimensional factor space.

Thus in practice we may wish to work with a subset of the available genes, particularly as the fitting of a mixture of factor analyzers will involve a considerable amount of computation time for an extremely large number of genes. Also, the intent of the cluster analysis may not be to produce a clustering of the tissues on the basis of all the available genes, but rather to discover and study different clusterings of the tissues corresponding to different subsets of the genes. Indeed, the simultaneous use of too many genes in the cluster analysis may serve only to create noise that masks the effect of a smaller number of genes. Therefore, the EMMIX-GENE program has two optional stages before the final stage of clustering the tissues. The first stage considers the selection of a subset of relevant genes from the available set of genes. The second stage then considers the grouping of the retained set of genes into a specified number (N_0) of groups. The third and final stage of the EMMIX-GENE approach concerns the clustering of the tissues by fitting mixtures of factor analyzers. It can be undertaken on the basis of (i) all or a selected subset of the available genes, (ii) all or some of the gene-group means, or (iii) all or some of the genes within a specified gene group.

4.1 Selection of relevant genes

We now describe the screening process used by EMMIX-GENE to select relevant genes for clustering the tissue samples into two clusters corresponding to, say, healthy and unhealthy tissues. This selection is undertaken in the absence of tissue samples that are of known classification with respect to the disease. The relevance of a gene for distinguishing between healthy and unhealthy tissue samples can be assessed on the basis of the value of $-2 \log \lambda$, where λ is the likelihood ratio statistic for testing $g = 1$ versus $g = 2$ components in the mixture model. In order to reduce the effect of atypically large observations on the value of λ, we fit mixtures of t components with their degrees of freedom inferred from the data. However, the use of t components in place of normal components still does not eliminate the effect of outliers on inference of the number of groups in the tissue samples. For example, suppose that for a given gene there is no genuine grouping in the tissues, but that there are a small number of gross outliers. Then a significantly large value of λ might be obtained, with one component representing the main body of the data (and providing robust estimates

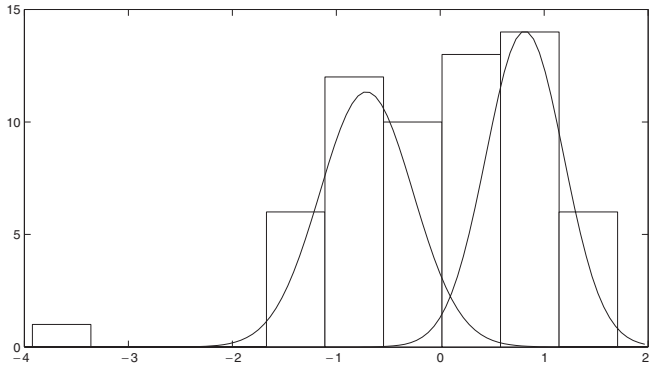

Fig. 1. Histogram of gene 1758 (H20819) with mixture of $g = 2$ fitted t components.

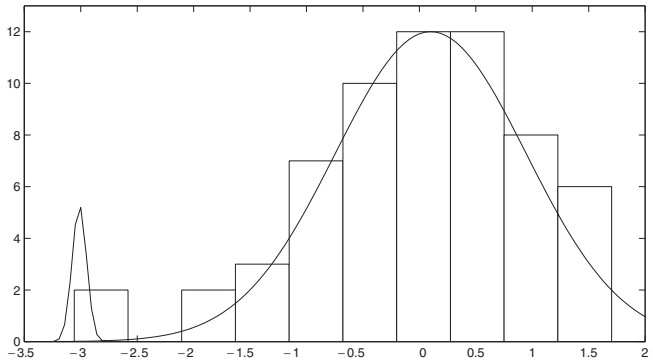

Fig. 2. Histogram of gene 474 (T70046) with mixture of $g = 2$ fitted t components.

of their underlying distribution) and the other representing the outliers. That is, although the t mixture model may provide robust estimates of the underlying distribution, it does not provide a robust assessment of the number of groups in the data.

Suppose now that for a given gene there are two groups in the tissue samples. If there are no outliers present in the tissue samples, we should obtain a significant value of λ with the two components of the fitted t mixture model corresponding to the two groups. But if there are outliers present, then the two components of the fitted t mixture model may still correspond to the two groups or it may happen that one component corresponds to the main body of the data and the other component to the outliers. An illustration of the former case is given in Figure 1 and of the latter case in Figures 2 and 3, using the data on two genes in 62 tissue samples from the colon cancer data of Alon *et al.* (1999).

In light of the above, the EMMIX-GENE software automatically assesses the relevance of each of the N

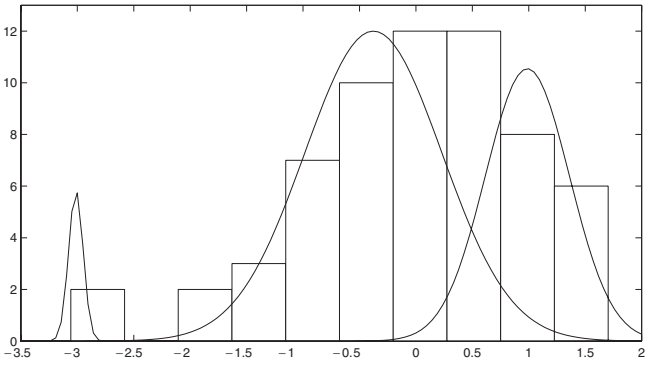

Fig. 3. Histogram of gene 474 (T70046) with mixture of $g = 3$ fitted t components.

genes by fitting one- and two-component t mixture models to the expression data over the M tissues for each gene considered individually. If $-2\log\lambda$ is greater than a specified threshold b_1,

$$-2\log\lambda > b_1 \tag{11}$$

then the gene is taken to be relevant provided that

$$s_{\min} \geqslant b_2, \tag{12}$$

where $s_{\min}$ is the minimum size of the two clusters implied by the two-component t mixture model and b_2 is a specified threshold.

If (11) holds but (12) does not for a given gene, then the three-component t mixture model is fitted to the tissue samples on this gene, and the value of $-2\log\lambda$ calculated for the test of $g = 2$ versus $g = 3$; see Figure 3. If (11) holds for this value of $-2\log\lambda$, the gene is selected as being relevant (provided at least two of the three clusters implied by the $g = 3$ solution have sizes not less than b_2). Although the null distribution of $-2\log\lambda$ for $g = 2$ versus $g = 3$ is not the same as for $g = 1$ versus $g = 2$ components, it would appear to be reasonable here to use the same threshold (11). The null distribution of $-2\log\lambda$ for the test of the null hypothesis $H_0 : g = g_0$ versus the alternative hypothesis $H_1 : g = g_1$ is unknown (for finite sample sizes) for normal or t components; see McLachlan and Peel (2000a, Chapter 6). Some simulations we performed for $g = 1$ versus $g = 2$ for t components suggest that the 90th percentile is around 9. In the examples to be discussed next, we took $b_1 = 8$. Concerning the lower bound on the minimum cluster size $s_{\min}$, we arbitrarily took $b_2 = 8$. In fitting the two- and three-component t mixture models to the tissue samples, we need to provide a starting point for the parameter estimate, or equivalently, the grouping of the data. This can be done by the user specifying a number

of random starts and a number of k-means-based starts. In our analyses to be presented later, we used four random and four k-means-based starts.

4.2 Grouping of genes

Concerning the end problem of clustering the tissue samples on the basis of the genes considered simultaneously, we could examine the univariate clusterings provided by each of the selected genes taken individually. But this would be rather tedious when a large number of genes have been selected. Thus with the EMMIX-GENE approach, there is a second (optional) stage for clustering the genes into a user-specified number (N_0) of groups by fitting a mixture of $g = N_0$ normal distributions with covariance matrices restricted to being equal to a multiple of the $(p \times p)$ identity matrix. That is, if the mixing proportions were fixed at 0.5, then it would be equivalent to using k-means and grouping the genes in terms of Euclidean distance between them. One could attempt to make a more objective choice of the number N_0 of groups by using, say, the likelihood ratio criterion or BIC. There is an extra complication here since the genes are not independently distributed within a tissue sample.

The groups of genes are ranked in terms of the likelihood ratio statistic calculated on the basis of the fitted mean of a group over the tissues for the test of a single versus two t components. A heat map of genes in a group versus the tissues is provided for each of the groups where, in each group, the tissues can be left in their original order or rearranged according to their cluster membership obtained by fitting a univariate t mixture model on the basis of the group mean. Alternatively, one could cluster the tissues by fitting a two-component mixture of factor analyzers on the basis of the genes within the group.

We have found in our analyses of microarray data sets that the means of the groups into which the genes have been clustered as above provide a useful representation of the genes in a lower dimensional space (the dimension of this space is equal to the number of groups N_0). If we cluster the tissues on the basis of the group means only, we are ignoring the relative sizes of the groups. This might have some impact on the accuracy of predictions if the aim were to construct a classifier for assigning the tissues to externally existing classes. For instance, one group may contain many genes that are useful in distinguishing between healthy and unhealthy. Thus if the genes within this group act independently, then there would be a loss in accuracy in using only the mean of this group and not making use of its size. But as the genes have been clustered into groups by working in terms of Euclidean distance (after normalization of the data), the impact of ignoring the size of the groups should be limited. This is because the genes within a group should in the main be at least moderately correlated with each other, as the Euclidean distance between any two genes is equal to $2(1-r)$, where r denotes the sample correlation between them.

5 IMPLEMENTATION

We illustrate the implementation of the EMMIX-GENE approach by applying it to two well-known data sets, the colon data of Alon *et al.* (1999) and the leukaemia data of Golub *et al.* (1999).

5.1 Clustering of colon tissues

Alon *et al.* used Affymetrix oligonucleotide arrays to monitor absolute measurements on expressions of over 6500 human gene expressions in 40 tumour and 22 normal colon tissue samples. These samples were taken from 40 different patients so that 22 patients supplied both a tumour and normal tissue sample. Alon *et al.* (1999) focussed on the 2000 genes with highest minimal intensity across the samples, and it is these 2000 genes that comprise our data set. The microarray data matrix $\boldsymbol{A}$ for this set thus has $N = 2000$ rows and $M = 62$ columns. In Alon *et al.* (1999), the tissues are not listed consecutively, but here we have rearranged the data so that the tumours are labelled 1–40 and the normals 41–62. Before we considered the clustering of this set, we processed the data by taking the (natural) logarithm of each expression level in $\boldsymbol{A}$. Then each column of this matrix was standardized to have mean zero and unit standard deviation. Finally, each row of the consequent matrix was standardized to have mean zero and unit standard deviation.

5.1.1 Clustering on basis of 446 genes. On the first stage of EMMIX-GENE, we selected 446 genes as relevant. It will be seen that the clustering of the tissue samples depends to a large extent as to which genes are selected for the feature variables to be used in the mixture model. In this sense, there may not be interest in attempting to find a clustering of the tissue samples on the basis of all the 2000 genes or even a reduced set such as the 446 genes deemed to be relevant. If there were still such interest, then one way to proceed is to fit a two-component mixture of factor analyzers to the tissues on the basis of, say, the 446 selected genes. We fitted mixtures of $g = 2$ factor analyzers for various levels of the number q of factors ranging from $q = 2$ to $q = 8$, but there was little difference between the clustering results. The clustering corresponding to the largest of the local maxima obtained gave the following clustering for $q = 6$ factors,

$$C_1 = \{1\text{–}12, 20, 25, 41\text{–}52\} \cup \{13\text{–}39, 21\text{–}24, 26\text{–}40, 53\text{–}62\}. \quad (13)$$

Getz *et al.* (2000) and Getz (2001) reported that there was a change in the protocol during the conduct of the microarray experiments. The 11 tumour tissue samples

(labelled 1–11 here) and 11 normal tissue samples (41–51) were taken from the first 11 patients using a poly detector, while the 29 tumour tissue samples (12–40) and normal tissue samples (52–62) were taken from the remaining 29 patients using total extraction of RNA. It can be seen from (13) that this clustering C_1 almost corresponds to the dichotomy between tissues obtained under the 'old' and 'new' protocols.

We also considered the clustering of the 62 tissue samples on the basis of the top 50 genes in the retained set of 446 genes. Fitting mixtures of factor analyzers with $q = 6$ factors, using 50 random and 50 k-means starts, we obtained the following clustering,

$$C_2 = \{1\text{–}26, 29, 31, 32, 34, 38, 41\text{–}52\} \cup \{27\text{–}28, 30, 33, 35\text{–}37, 39, 40, 53\text{–}62\}.$$

This clustering not only splits the tissue samples obtained under 'old' and 'new' protocols, but it also splits some of the 'new' tumour samples and some of the 'new' normal tissue samples.

5.1.2 Clustering on basis of gene groups. We now consider the clustering of the tissue samples after the retained set of 446 genes has been clustered into $N_0 = 20$ groups on the second stage of the EMMIX-GENE approach. A heat map of the genes in a group versus the tissues (and the heat map for the leukaemia data) may be viewed at http://www.maths.uq.edu.au/~gjm/emmix-gene/map.html. In Figure 4, we have plotted the 18 genes in the first group G_1 for the 62 tissues, with the latter arranged in order of the 40 tumours followed by the 22 normal tissues. In Figure 5, we give the corresponding plot of the 24 genes in the second group of genes G_2.

The clustering of the tissues on the basis of the 18 genes in G_1 using $q = 4$ factors in the mixture of factor analyzers model resulted in a partition C_3 of the tissues that is fairly similar to C_2, namely

$$C_3 = \{1\text{–}26, 29\text{–}32, 41\text{–}52, 55\text{–}56\} \cup \{27\text{–}28, 33\text{–}40, 53\text{–}54, 57\text{–}62\}.$$

The clustering of the tissues on the basis of the 24 genes in G_2 resulted in a partition of the tissues in which one cluster contains 37 tumours (1–29, 31–32, 34–35, 37–40) and 3 normals (48, 58, 60), and the other cluster contains 3 tumours (30, 33, 36) and 19 normals (41–47, 49–57, 59, 61–62). Calling this clustering C_4, we have that

$$C_4 = \{1\text{–}29, 31\text{–}32, 34\text{–}35, 37\text{–}40, 48, 58, 60\} \cup \{30, 33, 36, 41\text{–}47, 49\text{–}57, 59, 61\text{–}62\}.$$

It can be seen from Figure 4 that the clustering of the tissues on the basis of the genes in group G_1 gives two clusters with large intercluster differences between the tissues. The clusters are also quite cohesive, but this is accentuated by the fact that we are using genes that were put into the same group by carrying out the grouping effectively in terms of Euclidean distance between genes. Likewise, Figure 5 shows that the clustering of the tissues on the basis of the genes in group G_2 gives two cohesive clusters with a large intercluster differences. But it appears that the first clustering is stronger in terms of the likelihood ratio statistic λ formed from the individual genes in the groups and on their means. This clustering C_4 produced by the second group of genes G_2 is quite similar to the external classification, as its error rate is only 6.

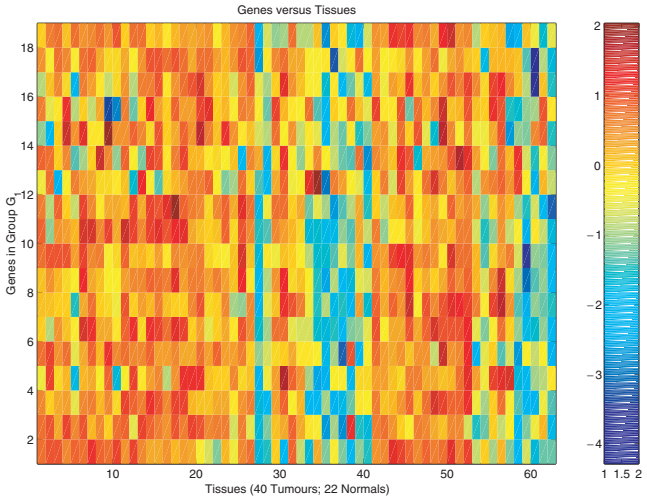

Fig. 4. Plot of 18 genes in group G_1 on the 40 tumour and 22 normal tissues.

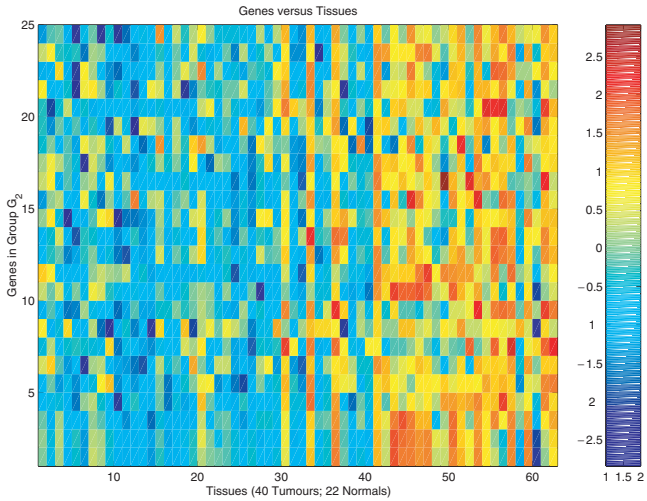

Fig. 5. Plot of 24 genes in group G_2 on the 40 tumour and 22 normal tissues.

It can be seen from Figure 5 that the genes in group G_2 tend to be more highly expressed in the normal tissues than in the tumours. Alon *et al.* (1999) and Ben-Dor *et al.* (2000) noted that the normal colon biopsy also included smooth muscle tissue from the colon walls. As a consequence, smooth muscle-related genes showed high expression levels in the normal tissue samples compared to the tumour samples, which generally had a low muscle content. Ben-Dor *et al.* (2000) identified a large number of muscle-specific genes as being characteristic of normal colon samples. We note that two of these genes (J02854 and T60155) are in group G_2, while group G_2 also contains two genes (M63391 and X74295) that Ben-Dor *et al.* (2000) suspected of being expressed in smooth muscle.

The six tissues that are misallocated under this second clustering (tumour tissues 30, 33, and 36 and normal tissues 48, 58, and 60) occur among those tissues that have been misallocated in other cluster and discriminant analyses of this data set. Tissues 30, 33, and 36 are taken from tumour tissue on patients labelled 30, 33, and 36 in Alon *et al.* (2000), while tissues 48, 58, and 60 are taken from normal tissue on patients 8, 34, and 36. These six tissues have been misallocated in previous analyses even in a discriminant analysis context where use is made of the external classification of these tissues. For example, with the support vector machine classifier formed in Chow *et al.* (2001) using the known classification of tissues, these six tissues along with tumour tissue 35 were misallocated in the (leave-one-out) cross-validation of this classifier. There is thus some doubt as to the validity of the so-called 'true' classification of these six tissues, which was determined by biopsy. An inspection of Figure 5 reveals that at least for the 24 genes in this plot, tumour tissues 30, 33 and 36 are very similar to the normal ones, while the normal tissues 48, 58, and 60 are very similar to the tumours. As explained in Chow *et al.* (2001), misclassification might be due to, say, simple error during sample handling, RNA preparation, data acquisition, and data analysis. They also noted that the normal tissues could have been misclassified because pathologically 'normal' regions of the colon could have substantial tumour-like properties from a molecular standpoint.

Applying a hierarchical procedure to cluster the 62 tissues on the basis of the 2000 genes, Alon *et al.* (1999) observed that the topmost division in the dendrogram divides the samples into two groups that misallocates three normal and five tumour tissues (tissues 2, 30, 33, 36, 37, 48, 52 and 58). The method used by Alon *et al.* (1999) can be viewed as fitting a normal mixture model with common spherical component–covariance matrices (although the variance was not estimated from the data; it was varied deterministically during the fitting process). Also, Alon *et al.* (1999) did not log the data. It is of interest to note that in fitting mixtures of diagonal normal components to the tissues on the basis of all the genes, the only way we could get the algorithm to converge to a local maximum that gave an implied clustering the same as C_4 or a perturbation of it (that is, similar to the external classification) was to use the unlogged data and to impose the condition of common spherical component–covariance matrices. Hence when the data are logged (as is appropriate), or when Euclidean distance is not used as the metric, the smooth muscle-related genes have a diminished capacity in the presence of other genes to distinguish between normal and tumour tissues.

5.1.3 Clustering on basis of group means. We also clustered the 62 tissues on the basis of the $N_0 = 20$ fitted group means obtained above by fitting a mixture of $g = 2$ factor analyzers for various levels of the number of factors q. The largest local maximum so located with $q = 8$ factors gives a clustering (C_5) that is similar to C_2 and C_3 with

$$C_5 = \{1\text{–}23, 25, 26, 41\text{–}52, 58\} \cup \{24, 27\text{–}40, 53\text{–}57, 59\text{–}62\}.$$

5.2 Leukaemia tissues

The EMMIX-GENE approach is applied now to the clustering of the leukaemia tissues of Golub *et al.* (1999), who studied gene expressions on two types of acute leukaemias: Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML). Gene expression levels were measured using Affymetrix high density oligonucleotide arrays containing $N = 7129$ genes on $M = 72$ tissues, comprising 47 cases of ALL (38 B-cell and 9 T-cell ALL) and 25 cases of AML. We have rearranged the order of the tissues so that the first 47 columns of the microarray data matrix $\boldsymbol{A}$ refer to the ALL cases and the next 25 to the AML cases. We followed the processing steps of Dudoit *et al.* (2001) of: (i) thresholding: floor of 100 and ceiling of 16 000; (ii) filtering: exclusion of genes with $\max / \min \leqslant 5$ and $(\max - \min) \leqslant 500$, where max and min refer respectively to the maximum and minimum expression levels of a particular gene across a tissue sample; (iii) the natural logarithm of the expression levels was taken (Dudoit *et al.*, 2001, used base 10 logarithms). This left us with 3731 genes. As with the normalization of the colon data in the previous example, we first standardized the columns of the matrix of the logged microarray data to have mean zero and unit standard deviation, and then we standardized the rows of this matrix to have mean zero and unit standard deviation. This preprocessing of the genes resulted in 3731 genes being retained.

We reduced this set further to 2015 genes by eliminating genes not considered to be relevant on the first stage of

the EMMIX-GENE approach. Proceeding to the second stage, we summarized the expression levels on these 2015 selected genes by clustering them into a number of groups ($N_0 = 40$). It was found that Groups 1 and 3 provide clusterings that are most similar to the external classification of the tissues. We subsequently confirmed this by fitting a two-component mixture factor analyzer with $q = 6$ factors to the tissues on the basis of the genes in Groups 1–3, respectively. The errors of allocation of the implied clustering corresponding to the largest local maximum located in each case were equal to 13, 35, and 6, respectively.

We also considered the clustering of the 72 tissue samples on the basis of the 40 fitted group means and the top fifty genes of the 2015 genes. To cluster the 72 tissues on the basis of the 40 group means, we fitted a mixture of $g = 2$ components with $q = 8$ factors. The local maximizer chosen from 50 random and 50 k-means-based starts gave a clustering with one tissue (number 69) misallocated.

We also fitted the same model to cluster the 72 tissues on the basis of the top fifty genes. Again using 50 random and 50 k-means-based starts, we obtained a clustering in which ten tissues were misallocated. This error dropped to one, when we started the mixture of factor analyzers from the true classification.

The 47 ALL tissues in these leukaemia data consist of 9 T-cell and 38 B-cell types. Given the existence of these three subclasses among the 72 tissues (25 AML, 9 T-cell ALL, and 38 B-cell ALL), Chow *et al.* (2001) considered the clustering of the 72 tissues into three groups. We decided to cluster the 72 tissues into three groups by fitting a three-component mixture factor analyzer with $q = 6$ factors. When this model was fitted from the 25–9–38 split of the tissues, it converged to a local maximizer that gives this split of the tissues apart from one B-cell ALL tissue that is put in the cluster corresponding to the T-cell ALL tissues. However, when we fitted the same model using 50 random and 50 k-means-based starts, we obtained a larger local maximum that gives a quite different split of the tissues into three clusters. One cluster consisted of the 25 AML cases plus 10 B-cell ALL cases; a second consisted of the 9 T-cell cases plus 5 B-cell ALL cases; the third cluster consisted of the remaining 23 B-cell ALL cases.

6 DISCUSSION

There has been increasing emphasis on a mixture model-based approach to clustering as it provides a sound mathematical-based method. However, in using this approach with mixtures of normal components that have nondiagonal covariance matrices, the number of observations to be clustered needs to be sufficiently large in number relative to their dimension in order to prevent singular estimates of the component–covariance matrices occurring during the estimation process. Unfortunately, this is not the case with the problem of clustering tissues on the basis of gene expression levels, as the latter are typically much larger than the number of tissues to be clustered. In this paper, we have shown how we can handle this clustering problem by adopting mixtures of factor analyzers to model the distribution of a high dimensional vector of gene expression data on a tissue. The proposed approach is demonstrated on two well known data sets in the microarray literature, the colon data of Alon *et al.* (1999) and the leukaemia of Golub *et al.* (1999). The aim was not to provide a detailed analysis of these sets, but rather to highlight the potential role and usefulness of a mixture model-based approach to the clustering of microarray expression data. In particular, we demonstrated how mixtures of factor analyzer models can identify various classes and subclasses among tissues on the basis of gene expression levels. Encouraging results are obtained in these two data sets for our proposed method for reducing the number of genes. For the leukaemia data set, the EMMIX-GENE approach yielded a two-cluster partition of the tissues that is consistent with the two types of acute leukaemia. However, for both data sets, we found clusterings of the tissues that do not correspond to the external (clinical) classification of the tissues, but do have an interpretation consistent with the biological background. For example, for the colon data, cluster analyses performed on the basis of various subsets of the genes selected as being relevant by EMMIX-GENE tended to provide strong support for a partitioning of the tissues into two classes that split the tissue samples obtained under 'old' and 'new' protocols. There is also support for the splitting of some of the 'new' tumour samples and some of the 'new' normal tissue samples, which can be partly explained by some of these tissues being outliers if the external classification is valid.

REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

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

Ben-Dor,A., Bruhn,L., Friedman,N., Nachman,I., Schummer,M. *et al.* (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–584.

Chow,M.L., Moler,E.J. and Mian,I.S. (2001) Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol. Genomics*, **5**, 99–111.

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

Dudoit,S., Fridlyand,J. and Speed,T.P. (2001) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, to appear.

Eisen,M.B., Spellmann,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Getz,G., Levine,E. and Domany,E. (2000) Coupled two-way clustering analysis of gene microarray data. *Cell Biol.*, **97**, 12 079–12 084.

Getz,G. (2001) Personal communication.

Ghahramani,Z. and Hinton,G.E. (1997) The EM algorithm for factor analyzers. *Technical Report No. CRG-TR-96-1*. The University of Toronto, Toronto.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gassenbeck,M. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hastie,T., Tibshirani,R., Eisen,M.B., Alizadeh,A., Levy,R., Staudt,L., Chan,W.C., Botstein,D. and Brown,P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, research0003.1–0003.21.

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. and Krishnan,T. (1997) *The EM Algorithm and Extensions*. Wiley, New York.

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

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

McLachlan,G.J., Peel,D., Basford,K.E. and Adams,P. (1999) The EMMIX software for the fitting of mixtures of normal and *t*-components. *J. Stat. Softw.*, **4**.

Meng,X.L. and van Dyk,D. (1997) The EM algorithm—an old folk song sung to a fast new tune (with discussion). *J. R. Stat. Soc.* B, **59**, 511–567.

Moler,E.J., Chow,M.L. and Mian,I.S. (2000) Analysis of molecular profile data using generative and discriminative methods. *Physiol. Genomics*, **4**, 109–126.

Nguyen,D.V. and Rocke,D.M. (2001) Tumor classification by partial least squares using microarray gene expression data. In *Methods of Microarray Data Analysis*. Kluwer, Dordrecht, pp. 109–124.

Xing,E.P. and Karp,R.M. (2001) CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, **17**, S306–S315.