On the classification of microarray gene-expression data

Kaye E. Basford, Geoffrey J. McLachlan and Suren I. Rathnayake

Submitted: 16th May 2012; Received (in revised form): 31st July 2012

Abstract
We consider the classification of microarray gene-expression data. First, attention is given to the supervised case, where the tissue samples are classified with respect to a number of predefined classes and the intent is to assign a new unclassified tissue to one of these classes. The problems of forming a classifier and estimating its error rate are addressed in the context of there being a relatively small number of observations (tissue samples) compared to the number of variables (that is, the genes, which can number in the tens of thousands). We then proceed to the unsupervised case and consider the clustering of the tissue samples and also the clustering of the gene profiles. Both problems can be viewed as being non-standard ones in statistics and we address some of the key issues involved. The focus is on the use of mixture models to effect the clustering for both problems.

Keywords: supervised classification; selection bias; unsupervised classification; mixture models; factor models; time-course data

INTRODUCTION
DNA microarray technology, first described in the mid-1990s, is a method to perform experiments on thousands of gene fragments in parallel. Its widespread use has led to a huge growth in the amount of expression data available. A variety of multivariate analysis methods such as cluster and discriminant analyses have been used to explore gene-expression data for relationships among the genes and the tissue samples. The utility of these methods has been demonstrated, for example, in the elucidation of unknown gene function, the validation of gene discoveries, the interpretation of biological processes, and the diagnosis and prediction of disease or treatment outcomes [1–3]. A common goal of microarray analyses of many diseases, in particular, cancer, is to identify as yet unclassified cancer subtypes for subsequent validation and prediction, and ultimately to develop individualized prognosis and therapy. Limiting factors include the difficulties of tissue acquisition and the expense of microarray experiments. Thus, often microarray studies attempt to perform an analysis of a small number of tumor samples on the basis of a large number of genes and can result in gene-to-sample ratios of ~100 fold. This is known as the 'big p, small n' problem in statistics, where p denotes the number of variables and n the number of observations. It is not the standard situation in statistics, where many procedures have been developed for the case where the variable-to-observation number ratio (corresponding to the gene-to-sample ratio) is relatively small.

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 M tissue samples), where we measure the expression levels of N genes in each experiment, the results can be represented by a $N \times M$ matrix. For each tissue, we can consider the expression levels of the...
N genes, called its ‘gene expression signature’. Conversely, for each gene, we can consider its expression levels across the different tissue samples, called its ‘gene expression profile’. The $M$ tissue samples might correspond to each of $M$ different patients or, say, to samples from a single patient taken at $M$ different time points. The $N \times M$ matrix is portrayed in Figure 1, where each sample represents a separate microarray experiment and generates a set of $N$ expression levels, one for each gene.

We firstly consider supervised classification, where the $M$ tissue samples are classified with respect to a number ($g$) of predefined classes (that is, their class of origin is known). The intent is to form a discriminant function (a classifier) on the basis of these classified $M$ tissue samples, $y_1, \ldots, y_M$, each consisting of the expression levels of $N$ genes for assigning a new unclassified tissue sample with gene signature $y_0$ to one of the $g$ classes. In this case, since the number of variables $p$ is equal to $N$, some form of variable selection usually is employed first or in conjunction with the formation of the classifier. For some classifiers, such as Fisher’s linear discriminant function [4], it is not possible to form the classifier before the number of genes is reduced to a level $p_0$, where is $p_0$ is less than $n - g$, where $n = M$ is the number of observations. In such situations, care has to be taken in the estimation of the error rate of a classifier formed from a reduced subset of genes that are selected in some optimal (non-random) way from the available genes. This is because there can be a selection bias [5].

We also consider unsupervised classification or clustering of the microarray data. Many researchers have explored the use of clustering techniques to arrange genes in some natural order, that is, to organize genes into clusters with similar behavior across relevant tissue samples (or cell lines). Although a cluster does not automatically correspond to a pathway, it is a reasonable approximation that genes in the same cluster have something to do with each other, such as being involved in the same pathway and/or having the same gene functions.

It can be seen that there are two distinct but related clustering problems with microarray data. One problem concerns the clustering of the tissues on the basis of the genes (the gene signatures) and the other concerns the clustering of the genes on the basis of the tissues (the gene profiles). This duality in cluster analysis is quite common. In the present context of microarray data, one may be interested in grouping tissues (patients) with similar expression values or in grouping genes on patients with similar types of tumors or similar survival rates.

**SUPERVISED CLASSIFICATION**

In this section, we consider the supervised classification of tissue samples. The intent is to form a classifier for assigning an unclassified sample with gene expression signature $y_0$ to one of $g$ classes. The classes may represent, for example, different types of a tumor, different types of treatment to be administered to a patient or the different outcomes (absence or presence of distant metastases within 5 years) for a patient undergoing a particular treatment such as chemotherapy. In this case, the $M$ available tissue samples represented in Figure 1 are of known classification with respect to the $g$ classes; that is, we know the class label $z_j$ corresponding to the $j$th gene signature vector $y_j$ ($j = 1, \ldots, n$), where $z_j = i$ if the $j$th tissue is from the $i$th class ($i = 1, \ldots, g; j = 1, \ldots, n$).

Because the dimension $p$ of each gene expression signature $y_j$ is very large, being equal to the number of genes $N$, some form of dimension or variable (gene) selection is usually undertaken first before an attempt is made to form a classifier from the training data. One common approach is to replace the $N$ genes by a smaller number $k$ of linear combinations of the genes; that is, the $N$ genes in each gene signature vector $y_j$ are replaced by the $k$ linear combinations $a_{1j}^T y_j, \ldots, a_{kj}^T y_j$. One way to choose the $k$ linear combinations is to perform a principal component analysis (PCA) and take them to be the first $k$ principal components of the sample covariance matrix. Another way is to use partial least squares

---

**Figure 1:** Gene expression data from $M$ microarray experiments represented as a matrix of expression levels with the $N$ rows corresponding to the $N$ genes and the $M$ columns to the $M$ tissue samples.
(PLS), whereby the \( a_1, \ldots, a_k \) are chosen by sequentially maximizing the covariance between the vector of class labels and the gene signatures. One would expect in general that PLS should perform better than PCA since it uses the known information on the class labels which are ignored with PCA; see Nguyen and Rocke [6].

There are several other methods of variable selection and formation of classifiers such as nearest-shrunken centroids [7]; see McLachlan et al. [8] for an account. There is much ongoing research in the area of variable selection in discriminant and cluster analyses in the case, where the dimension is much greater than the sample size [9].

In our example to follow, we adopted PLS to reduce the dimension of the gene signature to consist of \( k \) linear combinations of \( N_0 \) of the genes. Initially, the number \( N \) of available genes was reduced to \( N_0 \) by consideration of the variance of each gene across the \( M \) tissue samples. Then using these \( k \) linear combinations, we formed the support vector machine (SVM) as developed by Vapnik [10]. We used a linear kernel for the SVM classifier. Given the small number of samples \( n \), there is no gain to be had in using higher order kernels. Also, with linear kernels, Guyon et al. [11] have shown that the magnitude of the weight of a variable in the fitted SVM is proportional to its discriminatory worth. We used R statistical analysis software for the analysis and used the SVM implementation in the package e1071 with its default settings for a linear kernel.

Example 1: Pediatric leukemia gene expression data
To illustrate the aforementioned approach to the supervised classification of tissue samples, we consider the Pediatric Acute Lymphoblastic Leukemia (ALL) data of Yeoh et al. [12]. This data set had \( n=327 \) samples from nine subtypes of ALL and one normal group. The classes were BCR-ABL, E2A-PBX1, Hyperdip (>50), MLL, T-ALL, TEL-AML1, Hyperdip47-50, Hypodip, Normal and Pseudodip. We used here only the \( n=248 \) samples in the first \( g=6 \) classes. The final four classes were either too small to be reliably classified or there was probably a group structure present that was different from the given classes. There is some evidence for the last-mentioned possibility as Yeoh et al. [12] claim to have found a group within the last four classes that did not correspond to the given classes. The Affymetrix microarray software suite (MAS) version 4.0 software [13] was used in reading the raw data. The data consist of expression values as well as an indicator of the reliability of the readings. The reliability indicator denotes if the gene is present (reliable reading) or the reading is marginal or the gene is absent. We follow the approach of Yeoh et al. [12], where gene expression values for each sample were centered to have the same mean of 2500. For expression values labeled as present by MAS, values <1 were set to 1. Expression values that were not labeled as present by MAS were also set to 1. Genes, which had a range of values of <100 or which had less than 30 samples where that gene was identified as being present were deleted (Yeoh et al. [12] deleted genes that had <1% as being present, but we used a higher percentage to eliminate more potentially unrelated genes). This reduced the number of genes from 12625 to 6350. In order to further reduce the number of genes, we selected the top 1000 genes that have largest variance across the samples. The selected genes were then logarithmically transformed and then standardized to have mean 0 and unit variance.

In order to obtain an unbiased assessment of the true expected error rate of the SVM for a given \( k \) number of linear combinations of the genes obtained using PLS, we split the total set of tissue samples into two subsets of equal size. The SVM was trained on one of the subsets and tested on the other. This split was repeated randomly some 50 times and the final estimate of the true expected error rate was taken to be the mean-estimated error rate over these 50 splits.

This estimate of the true expected error rate is denoted by \( e_0 \) and is plotted in Figure 2 for the different values of \( k \), along with the corresponding values of \( e_A, e_1 \) and \( e_2 \). Here, \( e_A \) denotes the apparent error rate equal to the proportion of errors when the SVM is reapplied to the same subset on which it has

![Figure 2: Classification results of pediatric leukemia tumors using PLS.](image)
been formed. The estimated error rates $e_1$ and $e_2$ are the leave-one-out cross-validated rates with $e_1$ being formed by using the same $k$ linear combinations on each validation trial, whereas $e_2$ is formed by forming a new set of $k$ linear combinations on each validation trial, selected in the same manner as the original $k$ linear combinations were from the full training data set. Since a new set of $k$ linear combinations is not chosen on each validation trial in the formulation of $e_1$, it will provide too optimistic assessment of the error rate as pointed out in [5]. This is clearly evident in Figure 2, where it can be seen that $e_2$ is very close to the true error rate $e_0$, but $e_1$ is biased downward to the extent that it is 0 for values of $k > 9$.

The error rate, $e_2$, provides an almost unbiased estimate of the true expected error rate in the application of the classifier to new observation subsequent to the training data. Strictly speaking, we should on each validation trial in the formation of $e_2$, select a new set of genes (here $N_0 = 1000$) from the total number available, as was done in the original analysis. But as investigated in Zhu et al. [14], the consequence bias in not doing this is not of practical significance unless $N_0$ is relatively small, for example, a few hundred or less here.

Note also that if we were to take as our classifier the SVM with the smallest estimated error rate $e_2$ over the values of $k$ considered, then there would be a selection bias in using $e_2$ as its estimated error rate. We would need to perform an additional layer of cross-validation to remove this bias [15, 16].

**CLUSTERING OF TISSUE SAMPLES**

In contrast to discriminant analysis, in cluster analysis (unsupervised learning), there is no prior information on the group structure of the data or, in the case where it is known that the population consists of a number of classes, there are no data of known origin with respect to the classes. One of the difficulties of clustering is that the notion of a cluster is vague. A useful way to think about the different clustering procedures is in terms of the shape of the clusters produced. The majority of the existing clustering methods assume that a similarity measure or metric is known a priori; often the Euclidean metric is used. But clearly, it would be more appropriate to use a metric that depends on the shape of the clusters. The difficulty is that the shape of the clusters is not known until the clusters have been found and the clusters cannot be effectively identified unless the shapes are known.

Clustering methods can be categorized broadly as being hierarchical or non-hierarchical. With a method in the former category, every cluster obtained at any stage is a merger or split of clusters obtained at the previous stage. Hierarchical methods can be implemented in a so-called agglomerative manner (bottom–up), starting with $g = n$ clusters on a divisive manner (top–down), starting with $n$ entities to be clustered as a single cluster. In practice, divisive methods can be computationally prohibitive unless the sample size $n$ is very small. Hence hierarchical methods are usually implemented in an agglomerative manner.

One of the most popular non-hierarchical methods of clustering is $k$-means, where $k$ refers to the number of clusters to be imposed on the data. It aims to find $k = g$ clusters that minimize the sum of the squared Euclidean distance between each observation $y_i$ and its respective mean.

A common question in cluster analysis concerns the number of clusters in the data. This is a very difficult problem. Various criteria have been proposed; for example, the CH index [17], Hartigan’s index [18] and the gap statistic [19].

In the sequel, we focus on a model-based approach to clustering whereby a mixture model is adopted to represent the distribution of the feature vectors associated with the entities to be classified. We consider here the clustering of the $n = M$ tissue samples as represented by the gene signatures $y_1, \ldots, y_n$. In the next section, we consider the clustering of the gene profiles as represented by the rows of the data matrix in Figure 1.

With the mixture model approach to the clustering of the tissues on the basis of their associated gene signatures, each gene signature $y_j$ is assumed to have a $g$-component normal mixture density:

$$f(y_j; \Psi) = \sum_{i=1}^{g} \pi_i \phi(y_j; \mu_i, \Sigma_i),$$  \hspace{1cm} (1)

where $\phi(y_j; \mu_i, \Sigma_i)$ denotes the $p$-variate normal density function with mean $\mu_i$ and covariance matrix $\Sigma_i$ and the $\pi_i$ denote the mixing proportions, which are non-negative and sum to 1. Here, the vector $\Psi$ of unknown parameters consists of the mixing proportions $\pi_i$, the elements of the component means $\mu_i$ and the distinct elements of the component-covariance matrices $\Sigma_i$. It can be estimated by its maximum likelihood estimate $\hat{\Psi}$ calculated via the EM algorithm;
see [20, 21]. This approach gives a probabilistic clustering defined in terms of the estimated posterior probabilities of component membership \( \pi_i(y_j; \hat{\Psi}) \) \((i = 1, \ldots , g)\), where \( \pi_i(y_j; \hat{\Psi}) \) denotes the posterior probability that the \( j \)th feature vector with observed value \( y_j \) belongs to the \( i \)th component of the mixture \((i = 1, \ldots , g; j = 1, \ldots , n)\). Using Bayes’ theorem, it can be expressed as:

\[
\pi_i(y_j; \hat{\Psi}) = \frac{\pi_i \phi(y_j; \mu_i, \Sigma_i)}{\sum_{k=1}^{g} \pi_k \phi(y_j; \mu_k, \Sigma_k)}. \tag{2}
\]

It can be seen that with this approach, we can have a ‘soft’ clustering, whereby each observation may partly belong to more than one cluster. An outright clustering can be obtained by assigning \( y_j \) to the component to which it has the greatest estimated posterior probability of belonging. The number of components \( g \) in the normal mixture model can be assessed by consideration of the log likelihood by using, for example, criteria such as BIC [22].

As noted in [23], ‘clustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory’. Previously, Marriott [24, p. 70] had noted that the mixture likelihood-based approach ‘is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques and provides a test of significance for the results’. One potential drawback with this approach is that normality is assumed for the cluster distributions. However, this assumption would appear to be reasonable for the clustering of microarray data after appropriate normalization.

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or its more robust version in the form of the \( t \) density is that the implied clustering is invariant under affine transformations. Also, in the case where the components of the mixture correspond to externally defined subpopulations, the unknown parameter vector \( \Psi \) can be estimated consistently by a sequence of roots of the likelihood equation.

In a mixture model framework, the question of how many clusters there are in the data can be approached by consideration of the likelihood function. For example, by performing model selection using criteria such as Akaike’s AIC criterion or Schwarz’s BIC [22]. Another way is to use a resampling approach to test for the smallest value of \( g \) compatible with the data [25].

**Factor models**

The normal mixture model (1) cannot be directly fitted to the tissue samples if the number of genes \( p \) used in the expression signature is large. This is because the component-covariance matrices are highly parameterized with \( \frac{1}{2} p(p + 1) \) distinct elements each. A simple way of proceeding in the clustering of high-dimensional data would be to take the component-covariance matrices \( \Sigma_i \) to be diagonal. But this leads to clusters whose axes are aligned with those of the feature space, whereas in practice the clusters are of arbitrary orientation. For instance, taking the \( \Sigma_i \) to be a common multiple of the identity matrix leads to a soft-version of \( k \)-means which produces spherical clusters.

Banfield and Raftery [26] introduced a parameterization of the component-covariance matrix \( \Sigma_i \) based on a variant of the standard spectral decomposition of \( \Sigma_i (i = 1, \ldots , g) \). But if \( p \) is large relative to the sample size \( n \), it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it were possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when \( p \) is large relative to \( n \). Hence, in fitting normal mixture models with unrestricted component-covariance matrices to high-dimensional data, we need to consider first some form of dimension reduction and/or some form of regularization. A common approach to reducing the number of dimensions is to perform a PCA. However, the latter provides only a global linear model for the representation of the data in a lower-dimensional subspace. Thus it has limited scope in revealing group structure in a data set. A global non-linear approach can be obtained by postulating a finite mixture of linear (factor) submodels for the distribution of the full observation vector \( y_j \) given a relatively small number of (unobservable) factors [21, Chapter 8]. That is, we can provide a local dimensionality reduction method by a mixture of factor analyzers model, which is given by imposing on the component-covariance matrix \( \Sigma_i \) in Equation (1), the constraint:

\[
\Sigma_i = BB_i^T + D_i (i = 1, \ldots , g), \tag{3}
\]

where \( B_i \) is a \( p \times q \) matrix of factor loadings and \( D_i \) is a diagonal matrix \((i = 1, \ldots , g)\). We can think of the use of this mixture of factor analyzers model as being
purely a method of regularization. But, in the present context, it might be possible to make a case for it being a reasonable model for the correlation structure between the genes. This model implies that the latter can be explained by the linear dependence of the genes on a small number of latent (unobservable) variables specific to each component.

With the specification in Equation (3), the matrix of factor loadings $B_i$ is component-specific and so it requires $p \times q$ factors per component. To further reduce the number of parameters, Baek et al. [27] proposed mixtures of common factor analyzers (MCFA) whereby the matrix of factor loadings is common ($B_i$) to each component after transformation of the latent component factors to be white noise; that is, we take $B_i = BK_i$, $(i = 1, \ldots, g)$, where each $K_i$ is a $q \times q$ matrix.

With this additional assumption, the MCFA model is able to be fitted to data of dimension $p$ as large as 1000. However, in practice, we may wish to work with a subset of the available variables (genes in the present context), particularly as the fitting of the MCFA model will involve a considerable amount of computation time for an extremely large number $p$.

Indeed, the simultaneous use of too many variables (genes) in the cluster analysis may serve only to create noise that masks the effect of a smaller 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. For instance, the tissues (cell lines or biological samples) may cluster according to cell or tissue type (for example, cancerous or healthy) or according to cancer type (for example, breast cancer or melanoma). However, the same samples may cluster differently according to other cellular characteristics, such as progression through the cell cycle, drug metabolism, mutation, growth rate or interferon response, all of which have a genetic basis.

Therefore, the EMMIX-GENE procedure proposed by McLachlan et al. [28] has two optional steps before the final step of clustering the tissues. The first step considers the selection of a subset of relevant genes from the available set of genes by screening the genes on an individual basis to eliminate those which are of little use in clustering the tissue samples. The usefulness of a given gene to the clustering process can be assessed formally by a test of the null hypothesis that it has a single component normal or $t$-distribution over the tissue samples. The gene is retained if the test is significant at the specified level. Even after this step has been completed, there may still be too many genes remaining. Thus, there is a second step in EMMIX-GENE in which the retained gene profiles are clustered (after standardization) into a number of groups on the basis of Euclidean distance so that genes with similar profiles are put into the same group. In general, care has to be taken with the scaling of variables before clustering of the observations, as the nature of the variables can be intrinsically different. In the present context, the variables (gene expressions) are measured on the same scale. Also, as noted earlier, the clustering of the observations (tissues) via normal mixture models is invariant under changes in scale and location. The clustering of the tissue samples can be carried out on the basis of the groups considered individually using some or all of the genes within a group or collectively. For the latter, we can replace each group by a representative (a metagene) such as the sample mean as in the EMMIX-GENE procedure. The fitting process will be quite slow for such large values of $p$ and it is recommended that the dimension $p$ be reduced first to a smaller level.

The mixture of factor analyzers model is sensitive to outliers since it uses normal error and factors. Recently, Baek and McLachlan [29] have considered the use of mixture of $t$ analyzers in an attempt to make the model less sensitive to outliers.

**Example 2: Clustering of pediatric leukemia tumors**

To illustrate the mixture model-based approach, we now cluster the tumors in the ALL data of Yeoh et al. [12], which was analyzed in a supervised framework in the previous section. Starting with the 6350 genes as obtained by the filtering process adopted in the previous section, we further reduced the number of genes by running the select-gene step of the EMMIX-GENE procedure [28], whereby a gene is retained if twice the increase in the log likelihood is greater than a specified threshold in testing for a single $t$-component versus a mixture of two $t$-components. The null distribution for this likelihood ratio statistic for $t$-components in not known. Thus McLachlan et al. [28] determined a value for the threshold using a simulation study of the null distribution of this likelihood ratio statistic and selected the threshold to be 8. Following them, we also
selected the same threshold. Also, the minimum size of the two clusters is specified to be greater than an imposed threshold which avoids spurious clusters containing a few observations which are either outliers or lie in a lower dimensional subspace. We selected this threshold also to be 8 arbitrarily.

This reduced the data set to 3850 genes. The top 2000 genes in terms of the aforementioned likelihood ratio statistic were selected and clustered into 50 clusters using the cluster-genomes step of the EMMIX-GENE procedure, which uses essentially a soft-version of \( k \)-means to cluster the genes with the intent that highly correlated genes (genes close in Euclidean distance) are put together in the same cluster. We then took the means of these 50 clusters (metagenes) to be our \( p \) variables to which the MCFA approach was applied. Clustering of the gene profiles is discussed in detail in the next section. Here, we took 50 \( k \)-means clusters to provide groups of genes that are fairly similar to each other.

We implemented the MCFA approach with \( g = 6 \) components for the number of factors \( q \) ranging from 1 to 8. To measure the agreement between a clustering of the data and their true group membership, we calculated the error rate and the ARI (adjusted Rand index [30]). The results are presented in Table 1. We have also listed in this table under the heading of BIC, twice the negative of the log likelihood augmented by \( \log n \) times the number of (free) parameters \( d \). In a typical unsupervised clustering task, we do not know the true number of clusters in the data. In which case, one needs to select \( g \) and well as \( q \). For this particular example, by using \( g = 6 \), we demonstrate that MCFA is capable of revealing the true group structure in the data.

It can be seen that the maximum value of the ARI and the minimum error rate both occur for the same value 7 for the number of factors. In practice, in a clustering context, we do not know the value of these two indices of the goodness-of-fit of our MCFA model on which to choose. It can be seen that if we were to use BIC to choose \( q \) for the given number of components \( g \), we would not select the desired value here of \( q = 7 \); rather it would lead to the choice of \( q = 8 \). Hence work is ongoing on the development of criteria for the choice of the number of factors \( q \) with applications of the MCFA model. In unpublished work, the authors have obtained promising results for assessing the number of factors \( q \) via the stability of the clustering produced for different levels of \( q \).

To illustrate results that would be obtained using an hierarchical approach to clustering, we applied Ward’s agglomerative clustering algorithm [31]. When Euclidean distance was used, it gave for for \( g = 6 \) clusters an error rate of 0.33 (ARI = 0.42), while with the distance measure taken to be the centered Pearson dissimilarity, the corresponding error rate (ARI) was 0.18 (ARI = 0.74).

### Table 1: Clustering of pediatric leukemia tumors using MCFA

<table>
<thead>
<tr>
<th>( q )</th>
<th>BIC</th>
<th>ARI</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11190</td>
<td>0.00</td>
<td>0.73</td>
</tr>
<tr>
<td>2</td>
<td>9428</td>
<td>0.37</td>
<td>0.46</td>
</tr>
<tr>
<td>3</td>
<td>7815</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>4</td>
<td>6157</td>
<td>0.73</td>
<td>0.19</td>
</tr>
<tr>
<td>5</td>
<td>5292</td>
<td>0.87</td>
<td>0.12</td>
</tr>
<tr>
<td>6</td>
<td>4749</td>
<td>0.87</td>
<td>0.10</td>
</tr>
<tr>
<td>7</td>
<td>4666</td>
<td>0.91</td>
<td>0.07</td>
</tr>
<tr>
<td>8</td>
<td>4619</td>
<td>0.82</td>
<td>0.11</td>
</tr>
</tbody>
</table>

**CLUSTERING OF GENE PROFILES**

In this section, we consider the clustering of the gene profiles, which are represented by the rows of the data matrix depicted in Figure 1. In one sense, this problem would seem to be more straightforward than the problem of clustering the tumor samples on the basis of the gene signatures as covered in the previous section. This is because the dimension \( p \) of the gene profile vector is equal to the number \( M \) of tissue samples which is much smaller than the number \( n \) of sample points which is equal to the number of genes \( N \). However, with most applications of clustering as as the case with the clustering of the gene signatures, it is assumed that

(i) there are no replications on any particular entity specifically identified as such and

(ii) all the observations on the entities are independent of one another.

These assumptions should hold for the clustering of the tissue samples, although the tissue samples have been known to be correlated for different tissues due to flawed experimental conditions. However, condition (ii) will not hold for the clustering of gene profiles, since not all the genes are independently distributed and condition (i) will generally not hold either as the gene profiles may be measured over time or on technical replicates.
While this correlated structure can be incorporated into the normal mixture model (1) by appropriate specification of the component-covariance matrices $\Sigma$, it is difficult to fit the model under such specifications. For example, the M-step (the maximization step of the EM algorithm) may not exist in closed form. Accordingly, we now consider the EMMIX-WIRE model of Ng et al. [32], who adopt conditionally a mixture of linear mixed models to specify this correlation structure among the tissue samples and to allow for correlations among the genes. It also enables covariate information to be incorporated into the clustering process.

For $M$ microarray experiments as depicted in Figure 1, we have for the $j$th gene ($j = 1, \ldots, n$), where $n = N$, a feature vector (the gene profile vector), $y_j$, taken to be the $j$th row of the data matrix as depicted in Figure 1. The dimension $p$ of the profile vector $y_j$ is equal to the number of microarray experiments, $M$. Conditional on its membership of the $i$th component of the mixture, the EMMIX-WIRE procedure of Ng et al. [32] assumes that $y_j$ follows a linear mixed-effects model (LMM):

$$y_j = X\beta_i + Ub_i + Vc_i + \varepsilon_j (j = 1, \ldots, n),$$  

(4)

where the elements of $\beta_i$ (a $p$-dimensional vector) are fixed effects (unknown constants) ($i = 1, \ldots, g$). In Equation (4), $b_i$ (a $q_b$-dimensional vector) and $c_i$ (a $q_c$-dimensional vector) represent the unobservable gene- and tissue-specific random effects, respectively, conditional on membership of the $i$th cluster. These random effects represent the variation due to the heterogeneity of genes and tissues (corresponding to $b_i = (b_{i1}^T, \ldots, b_{ip}^T)^T$ and $c_i$, respectively). The random effects $b_i$ and $c_i$ and the measurement error vector $\varepsilon_j$ are assumed to be mutually independent. In Equation (4), $X$, $U$ and $V$ are known design matrices of the corresponding fixed or random effects. The dimensions $q_b$ and $q_c$ of the random effects terms $b_i$ and $c_i$ are determined by the design matrices $U$ and $V$ which, along with $X$ and $H$, specify the experimental design to be adopted.

With the LMM, the distributions of $b_i$ and $c_i$ are taken, respectively, to be multivariate normal $N_{q_b}(0, H_i)$ and $N_{q_c}(0, Q_i I_{q_c})$, where $H_i$ is a $q_b \times q_b$ covariance matrix and $I_{q_c}$ is the $q_c \times q_c$ identity matrices with dimensions being specified by the subscripts. The measurement error vector $\varepsilon_i$ is also taken to be multivariate normal $N_{q_e}(0, A_i)$, where $A_i = \text{diag}(W \xi_i)$ is a diagonal matrix constructed from the vector $(W \xi_i)$ with $\xi_i = (\sigma_{i1}^2, \ldots, \sigma_{iq_e}^2)^T$ and $W$ is a known $p \times q_e$ zero–one design matrix. That is, we allow the $i$th component-variance to be different among the $p$ microarray experiments.

The vector $\Psi$ of unknown parameters can be obtained by maximum likelihood via the EM algorithm, proceeding conditionally on the tissue-specific random effects $\zeta_i$. 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 $\hat{\zeta}_i$ ($i = 1, \ldots, g$).

This mixture model (4) of linear mixed distributions can be fitted using the program EMMIX-WIRE [32]. For example, it can be used to cluster time-course data where the genes are followed over time across a given number of tissues.

---

**Key Points**
- Classification of high-dimensional data sets such as microarray gene expression arrays is not trivial as the number of variables (the genes) is nearly always extremely large relative to the number of available samples (the tissue samples).
- The supervised classification problem concerning the prediction of the class labels of unclassified data points such as tissue samples is usually approached by first performing some form of dimension reduction such as partial least squares (PLS). The adopted classifier can then be formed on the basis of the gene expression signatures of reduced dimension. The assessment of the error rates of such classifiers will usually involve a selection bias of practical significance using standard methods of estimation such as ordinary cross-validation. Hence then cross-validation with additional layers of training and testing are needed on the validation trials.
- The unsupervised classification of observations such as tissue samples using a model-based approach has the challenge of being able to adequately represent the highly parameterized covariance matrices in the components of the mixture models adopted. One way of meeting this challenge is to adopt a factor-analytic representation of the component-covariance matrices. The clusters obtained by this approach in terms of available molecular data can be used to provide evidence for the existence of new subclasses of diseases and to explain why certain groups of patients with similar clinical characteristics respond differently to the treatment course.
- The clustering of genes as opposed to the tissue samples arises in various applications, including in time course studies where the behavior of the gene profiles over the tissues of interest can be approached by the fitting of mixtures of linear mixed models.

---

**FUNDING**

The research of GJ McLachlan and SI Rathnayake was supported by a grant from the Australian Research Council.
References