On the Simultaneous Use of Clinical and Microarray Expression Data in the Cluster Analysis of Tissue Samples

G.J. McLachlan, S, Chang, J. Mar, and C. Ambroise

Department of Mathematics
University of Queensland,
Department of Biostatistics Harvard University,
Centre National de la Recherche Scientifique
6599, 60200 Compiègne France
Contact: Geoff McLachlan, gjm@maths.uq.edu.au

Abstract

This paper considers a model-based approach to the clustering of tissue samples of a very large number of genes from microarray experiments. It 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. Frequently in practice, there are also clinical data available on those cases on which the tissue samples have been obtained. Here we investigate how to use the clinical data in conjunction with the microarray gene expression data to cluster the tissue samples. We propose two mixture model-based approaches in which the number of components in the mixture model corresponds to the number of clusters to be imposed on the tissue samples. One approach specifies the components of the mixture model to be the conditional distributions of the microarray data given the clinical data with the mixing proportions also conditioned on the latter data. Another takes the components of the mixture model to represent the joint distributions of the clinical and microarray data. The approaches are demonstrated on some breast cancer data, as studied recently in van't Veer et al. (2002).

Keywords: microarrays, gene expressions, mixture modelling, cluster analysis, clinical data

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), and Alizadeh et al. (2000), among others. In the past, mainly hierarchical methods have been applied to cluster analysis problems. However, attention is now turning to model-based approaches; see Ghosh and Chinnaiyan (2002), McLachlan et al. (2002), Pan et al. (2002), and Yeung et al. (2001).

The tissue space and the gene space are generally of quite different dimensionality ($10 \sim 10^2$ tissues versus $10^3 \sim 10^4$ genes). The clustering of the tissues on the basis of the genes is therefore a nonstandard problem in cluster analysis, as the dimension of each tissue sample (the number of genes) is so much greater than the number of tissues. One way to handle this dimensionality problem is to ignore the correlations between the genes and to cluster the tissue samples by fitting mixtures of normal component distributions with diagonal covariance matrices. This is essentially equivalent to using the k-means clustering procedure. However, this assumption of uncorrelated genes leads to spherical clusters whereas, in practice, the clusters tend to be elliptical due to the correlations that exist between some of the genes. This led McLachlan et al. (2002) to develop the software called EMMIX-GENE, which enables elliptical clusters to be imposed on the tissue samples. With a mixture model-based approach to clustering, the g components in the mixture model are conceptualized as representing the external classes corresponding to the g clusters to be imposed on the data. 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. 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 question of how many genuine clusters g are supported by the data can be considered in terms of the change in the likelihood function as additional components are included in the mixture model. The BIC criterion gives a guide as to how large the increase in the log likelihood must be for additional components to be included. A formal test can be constructed using a resampling approach as advocated in McLachlan (1987).

The EMMIX-GENE software is an extension of the EMMIX program, as developed by McLachlan et al. (1999) for standard clustering problems. The extension to the present case where the dimension of the feature vector (the number of genes) is so much greater than the number of cases to be clustered (the number of tissue samples) is handled in two ways. Firstly, the genes are screened on an individual basis to eliminate those which have little variation across all the tissue samples in terms of the likelihood ratio test statistic. Then the retained genes are clustered into groups on the basis of Euclidean distance so that highly correlated genes are clustered into the same group. The clustering of the tissue samples can then be carried out in terms of the group means. The groups of genes are ranked (in decreasing order of the clustering capacity of their means).

In practice, McLachlan et al. (2002) have found that typically only the means of the first dozen or so groups are needed to explore the tissue samples for any class and subclass structure. That is, the means of the groups into which the genes have been clustered provide a useful representation of the genes in a

lower dimensional space (the dimension of this space is equal to the number of groups). In the EMMIX-GENE software, a further reduction in the gene data can be obtained by fitting mixtures of factor analyzers to the group means. The use of the latter reduces the number of parameters by imposing the assumption that the correlations between the genes can be expressed in a lower space by the dependence of the tissues on a small number of (unobservable) factors.

In this paper, we consider the case where, in addition to the microarray expression data, there are also available data of a clinical nature on the cases on which the tissue samples have been recorded. The tissue samples can be clustered on the basis of the clinical and microarray data considered separately. But the simultaneous use of the clinical and microarray should lead to more powerful clustering procedures in situations where the clinical data contains information beyond that provided by the microarray experiments.

Two types of mixture models (unconditional and conditional) are proposed for the simultaneous use of clinical and microarray data for the clustering of tissue samples. With the unconditional approach, the mixture distribution models the joint distribution of the clinical and microarray data, while with the conditional approach, the mixture distribution models the conditional distribution of the microarray data given the clinical data.

These approaches are to be illustrated in the clustering of breast cancer tissues, as studied recently in van't Veer et al. (2002).

2 Mixture Models for Clinical and Microarray Data

It is supposed that the microarray data consist of n tissue samples, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, from n microarray experiments on p genes. that is, $\boldsymbol{y}_j$ is a p-dimensional vector. It is assumed further that there is available a vector $\boldsymbol{x}_j$ of clinical measurments taken on the jth case with tissue sample $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$. For the clustering of the n tissue samples (really the n cases) into g clusters, we shall fit a g-component mixture model, where the ith component represents the ith external class G_i corresponding to the ith cluster $(i = 1, \ldots, g)$. We let z_j be the (unobservable) class indicator associated with the jth tissue sample $\boldsymbol{y}_j$, where $z_j = i$ implies that the jth case is from the ith class $(i = 1, \ldots, g)$.

3 Unconditional Approach

The combined clinical and microarray data $(\boldsymbol{y}_j^T, \boldsymbol{x}_j^T)^T$ $(j = 1, \ldots, n)$ are taken to be n (independent) realizations from the mixture density,

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{x}) &= \sum_{i=1}^{g} \pi_i \, f_i(\boldsymbol{y}, \boldsymbol{x}) \\ &= \sum_{i=1}^{g} \pi_i \, f_i(\boldsymbol{x}) f_i(\boldsymbol{y} \mid \boldsymbol{x}), \end{aligned} \tag{1}$$

where $\pi_i = \mathrm{pr}\{Z = i\}$, $f_i(\boldsymbol{x})$ denotes the ith class-conditional density of the vector $\boldsymbol{x}$ of clinical features, and $f_i(\boldsymbol{y} \mid \boldsymbol{x})$ denotes the ith class-conditional density of the vector of the gene expression levels given the clinical-data vector $\boldsymbol{x}$ $(i = 1, \ldots, g)$. The symbol f is being used generically here to denote a density where, for discrete random variables, the density is really a probability function.

On specifying the forms of the densities of $f_i(\boldsymbol{x})$ and $f_i(\boldsymbol{y} \mid \boldsymbol{x})$, we can fit the mixture model (1) by maximum likelihood, using the EM algorithm of Dempster et al. (1977); see McLachlan and Krishnan (1997) and McLachlan and Peel (2002). In practice, the clinical features are usually nearly all discrete variables or are coded to be so. In discriminant and cluster analyses, it has been found that it is reasonable to proceed by treating discrete variables as if they are independently distributed within a class or cluster. This is known as the NAIVE assumption (Hand and Yi, 2001). Under this assumption, the ith class-conditional density of the vector of clinical features reduces to

$$f_i(\boldsymbol{x}) = \prod_{v=1}^{p} f_{iv}(x_v), \tag{2}$$

where $f_{iv}(x_v)$ denotes the ith class-conditional density of the vth clinical feature in $\boldsymbol{x}$.

Concerning the ith class-conditional density of the vector $\boldsymbol{y}$ of gene expressions given the clinical-data vector $\boldsymbol{x}$, we can take $\boldsymbol{y}$ not to depend on the clinical-data vector $\boldsymbol{x}$ and model its marginal density $f_i(\boldsymbol{y})$ by the multivariate normal density. For clinical features that are all discrete, we can allow for some dependence between the microarray-data vector $\boldsymbol{y}$ and the clinical-data vector $\boldsymbol{x}$ by adopting the location model as, for example, in Hunt and Jorgensen (1999). With the location model, $f_i(\boldsymbol{y} \mid \boldsymbol{x})$ is taken to be multivariate normal with a mean that is allowed to be different for some or all of the different levels of $\boldsymbol{x}$.

Given that the dimension p of the vector $\boldsymbol{y}$ of gene expressions is so much greater than the number n of available tissue samples, we would not be able to use all the genes in $\boldsymbol{x}$. In the examples to be presented later, we replace $\boldsymbol{x}$ by the vector of the means of the first 15 groups into which the genes have been clustered via the EMMIX-GENE software of McLachlan et al. (2002).

4 Conditional Approach

As an alternative to the use of the full mixture model (1), we may proceed conditionally on the realized values of the clinical-data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. This leads to the use of the conditional mixture model,

$$f(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{i=1}^{g} \pi_i(\boldsymbol{x}) f_i(\boldsymbol{y} \mid \boldsymbol{x}), \tag{3}$$

where $\pi_i(\boldsymbol{x})$ denotes the conditional probability that the class indicator takes on the value i given the vector $\boldsymbol{x}$ of clinical features. A common model for $\pi_i(\boldsymbol{x})$ is the logistic model under which

$$\pi_i(\boldsymbol{x}) = \frac{\exp(\beta_{i0} + \boldsymbol{\beta}_i^T \boldsymbol{x})}{1 + \sum_{h=1}^{g-1} \exp(\beta_{h0} + \boldsymbol{\beta}_h^T \boldsymbol{x})} \tag{4}$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^T$ for $i = 1, \ldots, g-1$, and

$$\pi_g(\boldsymbol{x}) = 1 - \sum_{h=1}^{g-1} \pi_h(\boldsymbol{x}).$$

5 Example on Breast Cancer Tissues

We illustrate the above approaches on some breast cancer data considered recently in van't Veer et al. (2002). In their paper, microarray experiments were performed on 98 primary breast cancers acquired from three classes of patients: 44 representing a good prognosis class (that is, those who remained metastasis free after a period of more than 5 years), 34 from a poor prognosis class (those who developed distant metastases within 5 years) and 20 representing a hereditary form of cancer, due to a BRCA1 (18 tumours) or BRCA2 (2 tumours) germline mutation.

Each microarray experiment involved an initial set of 24, 881 genes. To reduce the number of genes to something more computationally manageable, the same pre-processing filter used by van't Veer et al. (2002) was applied to the data at the outset of our analysis. The selection criteria of this filter required a gene to have both a P-value of less than 0.01 and at least a two fold difference in more than five out of the ninety-eight tissues for the gene to be retained. The initial set of genes was effectively reduced to 4,869 genes using this criterion.

The first step of the EMMIX-GENE algorithm was used to select the most relevant genes from this filtered set of 4,869 genes, reducing this number further to 1,867. The retained genes were then clustered into forty groups using the second step of the EMMIX-GENE algorithm, and the majority of gene groups produced were reasonably cohesive and distinct. The forty groups were ranked in decreasing order of the clustering capacity of their means. For the purposes of our illustration here, the vector $\boldsymbol{y}$ of microarray data was taken to be the means of the top 15 ranked groups. In Figures 1 and 2, we have displayed the expression levels for the genes in the first two groups obtained on the 78 tumours (44 without metastases followed by 34 with metastases), along with the 20 BRCA tumours.

The clinical data vector $\boldsymbol{x}$ consisted of six binary clinical variables, as considered in the Supplementary Information of van't Veer et al. (2002). The six variables comprised tumour grade ($x_1 = 0$, sizes 0 and 1, and $x_1 = 1$, size 2); oestrogen receptor (ER) status ($x_3 = 0, \leq 10$, and $x_3 = 1, > 10$); progesteron receptor (PR) status ($x_4 = 0, \leq 20$, and $x_4 = 1, > 20$); age ($x_5 = 0, \leq 40$, and $x_5 = 1, > 40$), and angioinvasion ($x_6 = 0$, no, and $x_6 = 1$, yes).

We clustered the $n = 78$ tissue samples into $g = 2$ clusters on the basis of the full (unconditional) model (1) and the conditional model (3) fitted to the microarray and clinical data together. We first report the results for the former model, using the NAIVE-independent version of it. That is, the binary variables are taken to be independent within a class, and the microarray-data vector $\boldsymbol{y}$ is taken to be independent of the clinical data (within a class).

The clustering corresponding to the largest of the local maxima located gave the following clustering:

$$\begin{aligned} C_1 \;=\; & \{1\text{–}17,\ 9\text{–}11,\ 13\text{–}23,\ 25\text{–}27,\ 29\text{–}42,\ 44\text{–}46\} \\ & \cup \{48\text{–}53,\ 56,\ 58\text{–}63,\ 66,\ 69\text{–}70,\ 72,\ 76\text{–}78\}. \end{aligned} \tag{5}$$

In Table 1, we have listed the fitted values of $f_{iv}(x_v = 1)$, which is the probability that the vth clinical (binary) variable is equal to one given its membership of the ith component of the mixture model ($i = 1, 2$). Concerning these clinical variables, it can be seen from Table 1 that the estimated probability of a high-grade tumour in the second component is close to one. However, high-grade tumours are not confined to this second component, as the estimated probability of a high-grade tumour in the first component is 0.426.

Table 1: Estimates of Component-Probabilities for Clinical Binary Variables

v	$f_{1v}(x_v = 1)$	$f_{2v}(x_v = 1)$
1 (grade)	0.426	0.968
2 (ER)	0.894	0.451
3 (PR)	0.787	0.226
4 (size)	0.277	0.710
5 (age)	0.745	0.710
6 (angioinvasion)	0.213	0.387

The first cluster contains 36 of the 44 tumours in the metastasis-free class G_1 of tumours; the second cluster, however, contains only 12 of the 34 tumours in the metastases class G_2. The misallocation rate of 22/34 for the metastases class G_2 is not surprising given the gene expressions as summarized in the groups of genes (see Figures 1 to 2). It can be seen from these heat maps that there are several tumours in class G_2 that have similar gene expression patterns as the tumours in class G_1. Thus it is very difficult to distinguish between the two tissue classes G_1 (metastasis-free) and G_2 (with metastases) on the basis of these gene expressions.

It would therefore be helpful if the clinical data could be used to aid in the cluster analysis of the tissue samples. But using just the microarray data (that is, ignoring the clinical data), we obtained the same clustering as C_1. Thus it would appear that for the purposes of clustering these tissue samples into two clusters corresponding to the external classes G_1 and G_2, the clinical data do not contribute any additional useful information.

We also fitted this mixture model by starting the EM algorithm from the external classification as given by the two classes G_1 and G_2. It led to a clustering corresponding to a smaller local maximum than C_1, where the first cluster still contains 36 of the 44 tumours in the metastasis-free class G_1, but where the second cluster corresponds to 22 of the 34 tumours in the metastases class G_2. This second clustering clearly corresponds more closely to the external calssification G_1 and G_2. But we could only locate it by starting the EM algorithm using this classification of the tissues. In any application of the EM algorithm using random starts, we always located the solution that yielded the first clustering C_1. It thus shows that this clustering C_1 corresponds to a local maximum of the likelihood function that dominates the other local maxima. Although C_1 does not correspond directly with the external classification G_1 and G_2, it is nonetheless an interesting finding. For example, it sheds light on the group structure that exists among the tissue samples in terms of the available clinical variables and gene expression data.

Concerning the use of the conditional model (3), we obtained similar results to those reported above for the full mixture model.

6 Supervised Classification

The misallocation rate of 22/34 for the second class G_2 of tumours with metastases is not surprising given the gene expressions as summarized in the groups of genes in Figures 1 and 2. Also, one has to bear in mind that in this cluster analysis, the tissues are classified in an unsupervised manner without using the knowledge of their true classification. But even when

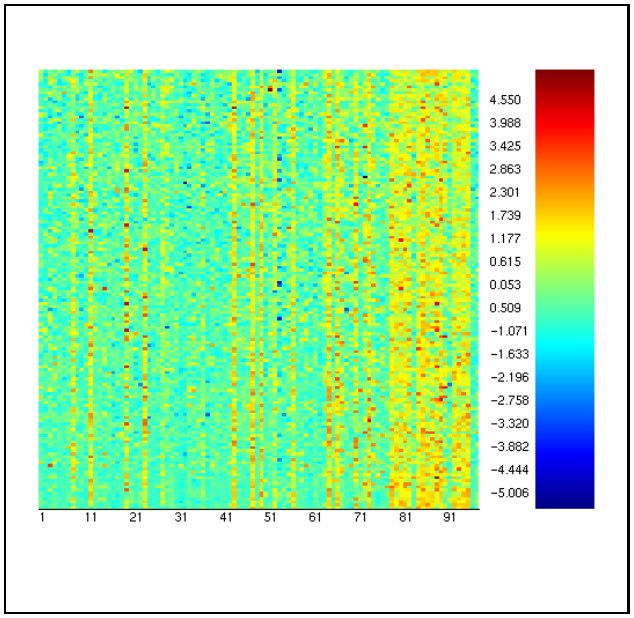

Figure 1: Plot of genes in the first ranked group of genes on 44 tissues in the disease-free class (G_1) followed by 34 tumours in the metastases class (G_2)

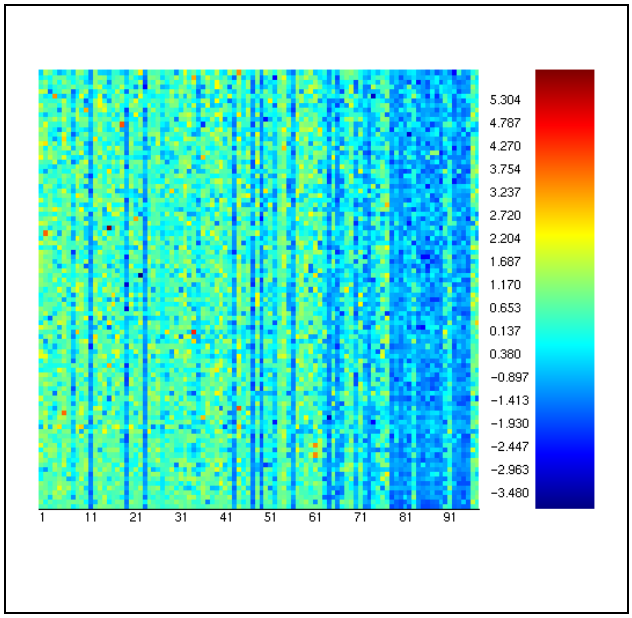

Figure 2: Plot of genes in the second ranked group of genes on 44 tissues in the metastasis-free class (G_1) followed by 34 tumours in the metastases class (G_2)

such knowledge was used (supervised classification) in van't Veer et al. (2002), the reported error rate was approximately 50% for members of G_2 when allowance was made for the selection bias in forming a classifier on the basis of an optimal subset of the genes (Ambroise and McLachlan 2002). Further analysis of this data set in a supervised context by Tibshirani and Efron (2002) confirmed the difficulty in trying to discriminate between the metastasis-free class and the metastases class.

We further investigated the predictive power of the microarray data when a rule is formed in a supervised manner (that is, with the knowledge of the class labels of the metastasis-free and and metastases classes, G_1 and G_2. A support vector machine (SVM) with a linear kernel was formed, as in Ambroise and McLachlan (2002). A recursive feature elimination (RFE) procedure was employed as in (Guyon et al., 2002), whereby a SVM was fitted, firstly, to all the genes, and then to reduced subsets of genes as they are eliminated progressively on the basis of the cross-validated error rate (CVAER). This rate is plotted in Figure 3, along with the apparent error rate (AER), and the estimated error rate FCV10AER obtained by performing a full 10-fold cross-validation, which also corrects for the selection bias incurred by selecting a subset of the genes in terms of the error rate; see Ambroise and McLachlan (2002) for further details. Most importantly, it can be seen from Figure 3 that when allowance is made for this selection bias, the estimated overall (error rate) of the SVM is quite high, its lowest value being just below 40% for a subset of genes of size between 2^5 and 2^7, approximately. Thus it is not surprising that the overall error rates of the clustering procedures above implemented without knowledge of the class labels have overall error rates approximately equal to 40%.

7 Discussion

The problem of clustering tissues on the basis of gene expression levels is not a straightforward problem, as the latter are typically much larger in number than the number of tissues to be clustered. There has been increasing emphasis on a mixture model-based approach to this problem, as such an approach provides a sound mathematical-based method. It also provides a framework for assessing whether any group structure located in the tissues is due to random fluctuations or represents a genuine grouping of the tissues. In this paper, we have proposed a mixture model-based approach in situations where there are also available clinical data recorded on the tissue samples to be clustered. We propose two mixture models: a full (unconditional) model and a conditional model formed in terms of the component-conditional distributions of the gene expression levels given the clinical data. With this latter model, the mixing proportions of the mixture model are modelled as a logistic function of the clinical variates.

Application of the full and conditional mixture models is illustrated on some breast cancer tumours as analyzed previously in van't Veer et al. (2002). These data consist of 78 tumours (a group G_1 of 44 metastasis-free tumours and a group G_2 of 34 tumours with metastases). The clustering implied by the maximum likelihood fit of the mixture model produces two clusters, one of which is strongly identified with the metastasis-free class of tumours G_1. However, only a minority of the class G_2 of 38 tumours with metastases are in the second cluster which would correspond to G_2 if the clustering were to represent this external classification.

As noted in the previous section, even when a prediction rule is formed in a supervised manner using the knowledge of the labels of the external classes G_1 and G_2, its (estimated) error rate is around 40%, which is not much less than the error rate of a rule that randomly assigns the tumours without attention to the microarray nor clinical data on a tumour. The clustering (unsupervised) and supervised results for this data set shows that approximately 60% of the tumours with metastases are similar in both the clinical variables and gene expression levels with the cluster of tumours that are metastasis-free after five years. Thus further biological research is required to produce clinical and microarray-based data that are capable of being able to differentiate breast cancer tu-

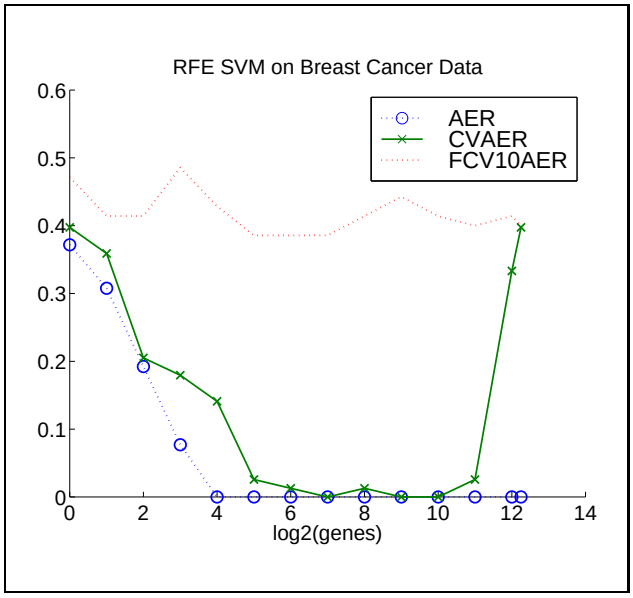

Figure 3: Estimated error rates of the SVM rule with RFE procedure

mours with good prognosis status from those of poor prognosis status.

References

Alizadeh, A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T.&Yu, X., et al. 2000, 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', *Nature* **403** 503–511.

Ambroise, C., &McLachlan, G.J. 2002 , Selection bias in gene extraction on basis of microarray gene expression data. 'Proceedings of the National Academy of Sciences USA' **99** 6562–6566.

Eisen, M.B., Spellmann, P.T., Brown, P.O., &Botstein, D. 1998, Cluster analysis and display of genome-wide expression patterns. 'Proceedings of the National Academy of Sciences USA' **95** 14863–14868.

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

Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D., &Brown, P.O. 1999, 'The transcriptional program in the response of human fibroblasts to serum', *Science* **283** 83–87.

McLachlan, G.J. 1987, 'On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture', *Applied Statistics* **36** 318-324.

McLachlan, G.J., Bean, R.W., and PeeL, D. 2002), 'A mixture model-based approach to the clustering of microarray expression data', *Bioinformatics* **18** 413–422.

Pan, W., Lin, J., and Le, C.T. 2002, 'Model-based cluster analysis of microarray gene expression data', *Genome Biology* *3*(2) research0009.1–0009.8.

Tibshirani, R.J., and Efron, B. 2002, 'Pre-validation and inference in microarrays. Statistical Applications in Genetics and Molecular Biology' **1** No. 1.

Van 'T Veer, L.J., Dai, H., Van de Vijver, M., He, Y.D., Hart, A.M., Mao, M., Peterse, H.L., Van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., &Friend, S.H. 2002, 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature* **415** 530–536.

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.