



Mixture modelling of gene expression data from microarray experiments

Debashis Ghosh¹ and Arul M. Chinnaiyan²

¹Department of Biostatistics and ²Department of Pathology, School of Public Health, University of Michigan, 1420 Washington Heights, Room M4057, Ann Arbor, MI 48109-2029, USA

Received on May 15, 2001; revised on August 17, 2001; accepted on September 18, 2001

ABSTRACT

Motivation: Hierarchical clustering is one of the major analytical tools for gene expression data from microarray experiments. A major problem in the interpretation of the output from these procedures is assessing the reliability of the clustering results. We address this issue by developing a mixture model-based approach for the analysis of microarray data. Within this framework, we present novel algorithms for clustering genes and samples. One of the byproducts of our method is a probabilistic measure for the number of true clusters in the data.

Results: The proposed methods are illustrated by application to microarray datasets from two cancer studies; one in which malignant melanoma is profiled (Bittner *et al.*, *Nature*, **406**, 536–540, 2000), and the other in which prostate cancer is profiled (Dhanasekaran *et al.*, 2001, submitted).

Availability: Macros written in the R language implementing the methods in this report can be obtained at the first author's website: <http://www.sph.umich.edu/~ghoshd/COMPBIO/mixture1/index.html>.

Contact: ghoshd@umich.edu

INTRODUCTION

DNA biochips have the potential of significantly impacting the study of human disease. By simultaneously gauging the expression of thousands of genes in clinical specimens, a wealth of data points is generated coalescing to form a molecular fingerprint of a disease process. As a result of the Human Genome Project, large collections of genes will be readily available for parallel genomic studies. Currently two platforms have emerged to dominate the microarray field—oligonucleotide arrays and spotted cDNA arrays. The first approach, developed by Affymetrix, involves *in situ* synthesis of oligonucleotides (less than 30 bases long) onto solid substrates using photolithographic techniques (Lipshutz *et al.*, 1999). These chips have been used for numerous applications including identification of Single Nucleotide Polymorphisms (SNPs) and monitoring of global gene expression.

For example, Golub *et al.* (1999) used oligonucleotide chips in the molecular classification of acute leukemias.

The second popular platform for studying global gene expression profiles was first pioneered by Patrick Brown and colleagues at Stanford and involves robotically printing cDNA clone inserts (200–2000 bp long) onto glass microscope slides (Brown and Botstein, 1999). These DNA chips are subsequently hybridized to different fluorescently labelled probes derived from RNA of various samples of interest. The power of this approach lies in its ability to comparatively analyze genome-wide patterns of mRNA expression relatively economically. Such experiments have been performed on lymphomas (Alizadeh *et al.*, 2000), breast cancers (Perou *et al.*, 2000) and cutaneous melanomas (Bittner *et al.*, 2000).

In the majority of studies using both technologies, the primary method of data analysis has been Hierarchical Clustering (HC; Hartigan, 1975; Everitt, 1993). This technique was introduced by Eisen *et al.* (1998) as a means to better visualize and interpret the high-dimensional data generated from cDNA microarrays. With HC, a matrix representing the pairwise dissimilarities between objects is first constructed using a distance metric; typically Euclidean distance is used. The objects can either be tissue samples or the individual genes. The algorithm begins with each object as a singleton cluster. The closest pair of clusters is found and merged. The dissimilarity matrix is then updated to take into account the merging; typically, the updating is done using the average-linkage method (Everitt, 1993). Based on this new dissimilarity matrix, the two closest distinct clusters are found and merged. This process is iterated until one cluster, which consists of all the samples, is left. HC has been utilized with some success in finding distinct subtypes of disease in several cancer studies (Alizadeh *et al.*, 2000; Bittner *et al.*, 2000).

As noted by Goldstein *et al.* (2001), HC has been utilized quite nondiscriminately in molecular profiling studies. Clustering methods are useful when the goal is to discover groupings in the gene expression data, and no external information exists (e.g. patient clinical

data). This has been referred to as class discovery in the literature (Golub *et al.*, 1999). A major issue with clustering techniques is the assessment of reliability of such algorithms. It is critical to be able to distinguish true clusters from those which arise by chance. This distinction is important to biologists who use microarray technology.

If one wishes to use HC, several methods exist for estimating the number of clusters in a dataset. The proposal of Calinski and Harabasz (1974) is to compute a suitably normalized ratio of between-and within-cluster sums of squares for a fixed number of clusters K . The value of K that maximizes this ratio is taken to be the estimate of the number of clusters. The method of Krzanowski and Lai (1985) is to compute K to maximize a difference quantity based on the within-cluster sums of squares. Neither of these methods allow for the case $K = 1$ (i.e. no clusters exist in the data). A method that can check for this scenario was proposed by Hartigan (1975). It is based on computing appropriately normalized ratios of between-cluster sums of squares and finding the smallest K for which the ratio is less than 10.

In this article, we explore the use of mixture model methodology for cDNA microarray data. Such techniques have been developed and utilized in other applications, such as brain imaging (Banfield and Raftery, 1993) and minefield detection (Dasgupta and Raftery, 1998). For cDNA microarray data, development of this methodology poses some unique challenges for which we develop some novel algorithms. While we consider data from cDNA microarray experiments here, similar considerations should hold for oligonucleotide microarray data as well (Lockhart *et al.*, 1996; Lipshutz *et al.*, 1999). An attractive feature of the mixture model approach is that it provides a statistical criterion for assessing the number of true clusters in the dataset of interest.

There have been recent proposals for joint probabilistic modelling of gene expression and sequence data. Barash and Friedman (2001) and Holmes and Bruno (2000) have proposed likelihood-based methods that iterate between clustering the gene expression data and looking for common regulatory motifs in the upstream regions of genes that are coordinately expressed. Our proposal differs from these in several respects. First, we do not consider the gene sequence data. In nonhomogeneous tissue from human cancers, looking for common regulatory motifs is problematic. We focus instead on modelling the gene expression data. Second, while the previous authors' methods apply only to clustering genes, we consider clustering genes and samples. Clustering samples is important in cancer studies because there is often interest in determining if gene expression profiles define molecular subtypes of disease. As we describe later, there is a fundamental asymmetry in the mixture model framework for clustering genes and samples. We develop methods in

order to make both types of clustering feasible.

SYSTEMS AND METHODS

Preprocessing the data

While not the focus of this article, a major issue in the analysis of microarray data is the preprocessing of the raw expression measurements. It might involve excluding expression ratios from certain genes and Expressed Sequence Tags (ESTs), subtracting a correction factor for background and/or multiplying by a constant in order to normalize the variation between chips. While some methods exist for the normalization of microarray data (Schuchhardt *et al.*, 2000; Yang *et al.*, 2001), we feel that there is no one uniformly best method and that the performance of any approach depends on the particular dataset. It will be assumed throughout that the data have already been normalized. In the examples in which the proposed techniques are applied, we will describe the preprocessing steps taken.

Model specification

Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ denote the observations to be clustered, where \mathbf{y}_i is a p -dimensional vector ($i = 1, \dots, n$). We will assume that the observed data are independent and identically distributed realizations from the density

$$f(\mathbf{y}_1, \dots, \mathbf{y}_n) \equiv \prod_{i=1}^n \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i | \mu_k, \Sigma_k), \quad (1)$$

where π_k ($k = 1, \dots, K$) is the probability that an observation belongs to the k th group, and

$$f_k(\mathbf{y}_i | \mu_k, \Sigma_k) \equiv |2\pi \Sigma_k|^{-p/2} \exp\{-(\mathbf{y}_i - \mu_k)^T \times \Sigma_k^{-1} (\mathbf{y}_i - \mu_k)\} \quad (2)$$

is a multivariate normal density with mean μ_k and covariance matrix Σ_k . Thus in (1), we have formulated that the gene expression data arise from a mixture of K multivariate normal populations. While the assumption of normality might seem restrictive at first, mixtures of normal distributions yield a variety of distributions and are quite flexible in practice. The distributions of the K components are fully specified by μ_k and Σ_k , $k = 1, \dots, K$. In the current framework, these are left unspecified. Banfield and Raftery (1993) proposed the following eigenvalue parametrization for Σ_k ($k = 1, \dots, K$):

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T, \quad (3)$$

where \mathbf{D}_k is the matrix of eigenvectors of Σ_k , λ_k is a constant of proportionality, and \mathbf{A}_k is the diagonal matrix of values proportionate to the eigenvalues of Σ_k . There is a nice geometric interpretation implied by (3). The parameter λ_k controls the volume for the k th component,

Table 1. Possible covariance structures for Σ_k ($k = 1, \dots, K$) for agglomerative HC and EM algorithm

Variance model	HC	EM	Distribution	Volume	Shape	Orientation
$\lambda \mathbf{I}$	X	X	Spherical	Equal	Equal	NA
$\lambda_k \mathbf{I}$	X	X	Spherical	Variable	Equal	NA
$\lambda \mathbf{DAD}^T$	X	X	Ellipsoidal	Equal	Equal	Equal
$\lambda \mathbf{D}_k \mathbf{AD}_k^T$		X	Ellipsoidal	Equal	Equal	Variable
$\lambda \mathbf{D}_k \widehat{\mathbf{A}} \mathbf{D}_k^T$	X		Ellipsoidal	Equal	Fixed	Variable
$\lambda_k \mathbf{D}_k \mathbf{AD}_k^T$		X	Ellipsoidal	Variable	Equal	Variable
$\lambda_k \mathbf{D}_k \widehat{\mathbf{A}} \mathbf{D}_k^T$	X		Ellipsoidal	Variable	Fixed	Variable
$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$	X	X	Ellipsoidal	Variable	Variable	Variable

\mathbf{A}_k its shape and \mathbf{D}_k its orientation. These characteristics can be estimated from the data. While we have placed no cluster-level constraints on shape, orientation or volume in (3), it is possible to have some of these parameters take the same values across clusters. We discuss this issue further in the next section.

ALGORITHM

There are two steps in fitting model (1) to the data. Maximum likelihood estimation of the parameters in the mixture model are fit using the Expectation–Maximization (EM) algorithm (Dempster *et al.*, 1977). However, the performance of the EM algorithm relies critically on the initial values. We compute these values by model-based hierarchical agglomerative clustering, which we now describe.

Initialization via hierarchical agglomerative clustering

Suppose that for each observation \mathbf{y}_i ($i = 1, \dots, n$), the true cluster membership $l_i \in \{1, \dots, K\}$ is known. We can then define a classification log-likelihood

$$l^{CL}(\theta_1, \dots, \theta_K; l_1, \dots, l_n | \mathbf{y}_1, \dots, \mathbf{y}_n) \equiv \prod_{i=1}^n f_{l_i}(\mathbf{y}_i | \theta_{l_i}), \quad (4)$$

where $\theta_k = (\mu_k, \Sigma_k)$ and $f_{l_i}(\mathbf{y}_i | \theta_{l_i})$ is the density equation (2) ($i = 1, \dots, n; k = 1, \dots, K$). Potential specifications for Σ_k ($k = 1, \dots, K$) can be found in Table 1.

In our situation, however, the true cluster memberships are unknown. We use agglomerative HC in order to find a maximizer of (4). This procedure is similar in spirit to other HC algorithms; it starts with each observation as a singleton cluster. All possible pairs of clusters are considered, and the pair that leads to the greatest increase in (4) is then merged. This procedure is iterated repeatedly until one cluster remains. By default, this cluster will

have all n observations. Note that with agglomerative HC, the clusters are merged according to the probabilistic criterion (4). This is different from ordinary HC methods, where the merging of clusters is based on a combination of the dissimilarity matrix and a method of defining distance between clusters (e.g. average linkage). However, it can be shown that the two classes of methods are equivalent under certain situations. For example, the method of Ward (1963) can be derived as a special case of model-based agglomerative clustering with $\Sigma_k = \lambda \mathbf{I}$ in (4).

The output of hierarchical agglomerative clustering is the assignment of observations to clusters, along with the estimated mean and covariance for each cluster. This serves as starting values for maximum likelihood estimation of the parameters in (1). The estimation will be done using the EM algorithm (see Section Expectation–Maximization algorithm).

Expectation–Maximization algorithm

In the EM algorithm framework, it is useful to phrase the maximum likelihood estimation problem as a missing data problem. For this setting, the complete data are $\mathbf{x}_i = (\mathbf{y}_i, \mathbf{z}_i)$ ($i = 1, \dots, n$), where $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ is defined by

$$z_{ik} = \begin{cases} 1 & \text{if } \mathbf{y}_i \text{ belongs to group } k \\ 0 & \text{otherwise.} \end{cases}$$

We assume that \mathbf{z}_i , $i = 1, \dots, n$, are iid realizations from a multinomial distribution with probabilities π_1, \dots, π_K ($\sum_{k=1}^K \pi_k = 1$) and that the density of \mathbf{y}_i given \mathbf{z}_i is $\prod_{k=1}^K f_k(\mathbf{y}_i | \theta_k)^{z_{ik}}$. Then it is easy to then derive the complete-data likelihood and log likelihood:

$$L^{CD}(\theta_1, \dots, \theta_K, \pi_1, \dots, \pi_K | \mathbf{x}_1, \dots, \mathbf{x}_n) \\ = \prod_{i=1}^n \prod_{k=1}^K \{\pi_k f_k(\mathbf{y}_i | \theta_k)\}^{z_{ik}},$$

and

$$l^{CD}(\theta_1, \dots, \theta_K, \pi_1, \dots, \pi_K | \mathbf{x}_1, \dots, \mathbf{x}_n) \\ = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log\{\pi_k f_k(\mathbf{y}_i | \theta_k)\}. \quad (5)$$

Again, we must specify a model for Σ_k ($k = 1, \dots, K$); a list of potential models can be found in Table 1. The missing data here are the \mathbf{z}_i ($i = 1, \dots, n$), which represent the cluster assignments. Given estimates $\hat{\pi}_1, \dots, \hat{\pi}_K$ and $\hat{\theta}_1, \dots, \hat{\theta}_K$, the E-step of the EM algorithm involves estimating \mathbf{z}_i by $E[\mathbf{z}_i | \mathbf{y}_i, \hat{\pi}_1, \dots, \hat{\pi}_K, \hat{\theta}_1, \dots, \hat{\theta}_K]$. The estimator here has a simple form:

$$\hat{z}_{ik} = \frac{\hat{\pi}_k f_k(\mathbf{y}_i | \hat{\theta}_k)}{\sum_{j=1}^K \hat{\pi}_j f_j(\mathbf{y}_i | \hat{\theta}_j)} \quad (i = 1, \dots, n; k = 1, \dots, K).$$

The estimated \mathbf{z}_i ($i = 1, \dots, n$) are then plugged into (5), and the complete data log-likelihood is then maximized as a function of θ_k and π_k , $k = 1, \dots, K$. This is the M-step of the EM algorithm. Estimates of π_k and θ_k ($k = 1, \dots, K$) are output from the M-step and are then input into the E-step. The two steps (E-step and M-step) are then iterated until convergence is reached. Many authors have shown (Wu, 1983; Boyles, 1983) that under general regularity conditions, the solution from the EM algorithm will converge to a local maximum. In practice, the results from fitting the algorithm has proven to be acceptable. One of the potential problems with the EM algorithm is its rate of convergence in practice. It has converged rapidly in the examples we have analyzed.

Selecting the number of clusters

In the previous sections, it was implicitly assumed that the number of clusters, K , was fixed. However, one of the questions of scientific interest is assessing the reliability of the output from their clustering analyses. This is equivalent to the question of determining the number of true clusters that exist in the data and for determining the number of components in a mixture model (Roeder and Wasserman, 1997). It is hard to apply classical statistical hypothesis testing methods because the usual regularity conditions are not satisfied. In this section, we describe a framework for determining the number of clusters based on Bayes factors (Kass and Raftery, 1995).

The Bayes factor for determining whether there are k clusters in the data versus l clusters is given by

$$B_{kl} = \frac{f(\mathbf{y}|K = k)}{f(\mathbf{y}|K = l)}, \quad (6)$$

where $f(\mathbf{y}|K = k) = \int f(\mathbf{y}|K = k, \theta_k) p(\theta_k) d\theta_k$, and $p(\theta_k)$ is the prior density for θ_k . The integration is done over the support of $p(\theta_k)$. A similar expression holds for $f(\mathbf{y}|K = l)$. If $B_{kl} > 1$, then the data are providing some evidence for k clusters against l clusters. Values of $B_{kl} > 100$ provide very strong evidence for k clusters relative l clusters in the data.

There are several advantages to the use of Bayes factors for determining the number of clusters. First, the approach does not rely on asymptotic theory, unlike most classical hypothesis testing procedures. In addition, for this situation we are typically comparing nonnested models. While this comparison cannot be performed using usual hypothesis testing methods, Bayes factors allow for the comparison of nonnested models.

The major problem with using (6) is that it involves evaluation of two integrals. With large datasets, direct evaluation of the Bayes factor is not practically feasible. To address this problem, we will utilize an approximation. We first assume that all models have an equal *a priori*

probability. Then, an approximation for (6) based on the Bayesian Information Criterion (BIC, Schwarz, 1978) can be computed:

$$\begin{aligned} 2 \log B_{kl} &= 2 \log f(\mathbf{y}|K = k) - 2 \log f(\mathbf{y}|K = l) \\ &\approx 2 \log f(\mathbf{y}|\hat{\theta}_k, K = k) - v_k \log n \{2 \log f \\ &\quad \times (\mathbf{y}|\hat{\theta}_l, K = l) - v_l \log n\} \\ &\equiv \text{BIC}_k - \text{BIC}_l, \end{aligned}$$

where v_j is the number of independent parameters to be estimated in the j -component mixture model, and BIC_j is the corresponding value for the Bayesian Information Criterion. Note that the number of independent parameters will depend on what type of covariance model is assumed for (3). This approximate Bayes factor has tended to work well in applications in minefield detection and astrophysics (Dasgupta and Raftery, 1998; Campbell *et al.*, 1999). In addition, in some simulation studies we performed (data not shown), we found the approximation to work well.

IMPLEMENTATION

Mixture modelling in microarray studies

There are certain aspects of experiments involving microarray data which make direct application of mixture models impractical. First, thousands of genes and ESTs are typically considered for analysis. As was mentioned earlier, the hierarchical agglomerative procedure begins with each observation as a singleton cluster. If we wish to cluster genes in this manner, this requires computational storage that is quadratic in the number of genes considered. For most practical situations, this is still not feasible.

Another feature of microarray studies is that the number of samples profiled is typically much smaller than the number of genes on a microarray. Thus, if we wish to cluster samples, we are in a situation where p is much larger than n . It is impossible to fit model (1) to these data.

Thus, there is an asymmetry in trying to apply mixture models for clustering in microarray studies. If we wish to cluster genes, then there is sufficient data for fitting the mixture model, but it is not computationally feasible. If we want to cluster samples, then it is not possible to fit (1) with the raw data because of the relative magnitudes of p and n .

In the next two sections, we describe algorithms for utilizing mixture models in order to cluster genes and samples.

Mixture models for clustering of genes

Before using model-based agglomerative hierarchical clustering, we perform an initial clustering of the data. The goal of this clustering is to reduce the number of clusters initially considered so that the computation becomes

more manageable. This is simply a preprocessing step. In practice, we group the original set of genes into $K = 1000$ clusters of genes. These cluster assignments are then used as the input for model-based agglomerative HC. We have typically used k -means clustering (MacQueen, 1967) for the preprocessing step; however, any other type of partition clustering method could be used. The effect of varying both K and the type of partition clustering algorithm on the mixture model output is examined for the prostate cancer data (Dhanasekaran *et al.*, 2001) in Section **Results**.

Because the number of genes is large, it is possible that there are several possible maxima. We recommend two strategies to address this issue. First, the EM algorithm should be initialized using several different starting values in order to diagnose the sensitivity of the results. Second, we examine the list of genes that are clustered using the EM algorithm to see if the clustering makes biological sense.

Mixture models for clustering of samples

As mentioned earlier, it is obvious that the dimension of the genes, p , must be reduced in order to fit the mixture model given in (1). We have employed principal components analysis (Anderson, 1984, Chapter 11) for dimension reduction. With this method, the linear combinations of the genes with maximal variance that are uncorrelated with each other are found.

There are several aspects of principal components analysis which make its use desirable in this context. First, the dimension reduction can be performed in a totally unsupervised manner. Second, because principal components analysis transforms data using a linear combination of the original variables, it can be shown that the transformed data also has a mixture model structure similar to that in (1). Third, the mixture model can now be fit to the transformed data using a subset of the principal components. It should be noted that our primary goal with this analysis is to determine the number of clusters that exist in the set of samples; we do this using Bayes factors. We are not attempting to interpret the principal components themselves. Discussion of the application of principal components techniques to actual datasets is given in Sections **Results** and **Discussion**.

Software

There are many software packages available for fitting mixture models. We have chosen to utilize the package MCLUST (Fraley and Raftery, 1999), which is available for the public domain statistical software package R (<http://www.r-project.org/>). MCLUST can be found at the following website: <http://www.stat.washington.edu/fraley/mclust/soft.shtml>.

We have written two macros in R for implementing

the mixture model-based clustering methods we have proposed for genes and samples. They are obtainable from the first author's website at the following URL: <http://www.sph.umich.edu/~ghoshd/COMPBIO/mixture1/index.html>.

RESULTS

Cutaneous melanoma data

In this section, we consider data from the cDNA microarray experiments performed by Bittner *et al.* (2000). For the analysis presented here, we consider the 31 melanoma samples. The microarray used in this study contained 8150 human cDNAs, 6912 of which were reported to be sequence verified. Of these cDNAs, 3613 were included in the analysis based on having average mean expression levels above background across all experiments greater than 2000 arbitrary units for the least intense signal (Cy3 or Cy5), and having average spot sizes greater than 30 pixels for all experiments.

Expression ratios of Cy5/Cy3, were calculated for the 3613 well-measured genes. The following normalization steps were taken (Radmacher, Personal communication):

- (1) ratios greater than 50 and less than 0.02 were truncated to 50 and 0.02, respectively;
- (2) the resulting ratios were transformed to a logarithm scale (base 2);
- (3) the log-ratios were normalized by subtracting the median log-ratio within an experiment (i.e. within a slide) from all log-ratios for that experiment. This results in the median log-ratio within an experiment being zero.

It should be pointed out that no normalization was performed across experiments, since a single reference probe was used for all experiments.

In their paper, Bittner *et al.* (2000) used an average-linkage HC procedure for clustering samples. A replica of their dendrogram can be found in Figure 1. Based on cutting the dendrogram at a value of 0.54, they found a cluster of 19 melanomas. Subsequently, they developed some tests for assessing the reliability of this cluster. Before applying our proposed clustering procedures, we used the methods for estimating the numbers of clusters described in Section **Introduction**. The procedures of Calinski and Harabasz (1974) and Hartigan (1975) yield a value of two clusters, implying that the dendrogram in Figure 1 should be truncated at a value between 1.05 and 1.15. The method of Krzanowski and Lai (1985) yields a value of eight clusters, implying the same cutoff value as that used by Bittner *et al.* (2000).

Next, we now apply the mixture methodology for clustering samples (i.e. the clinical melanoma specimens).

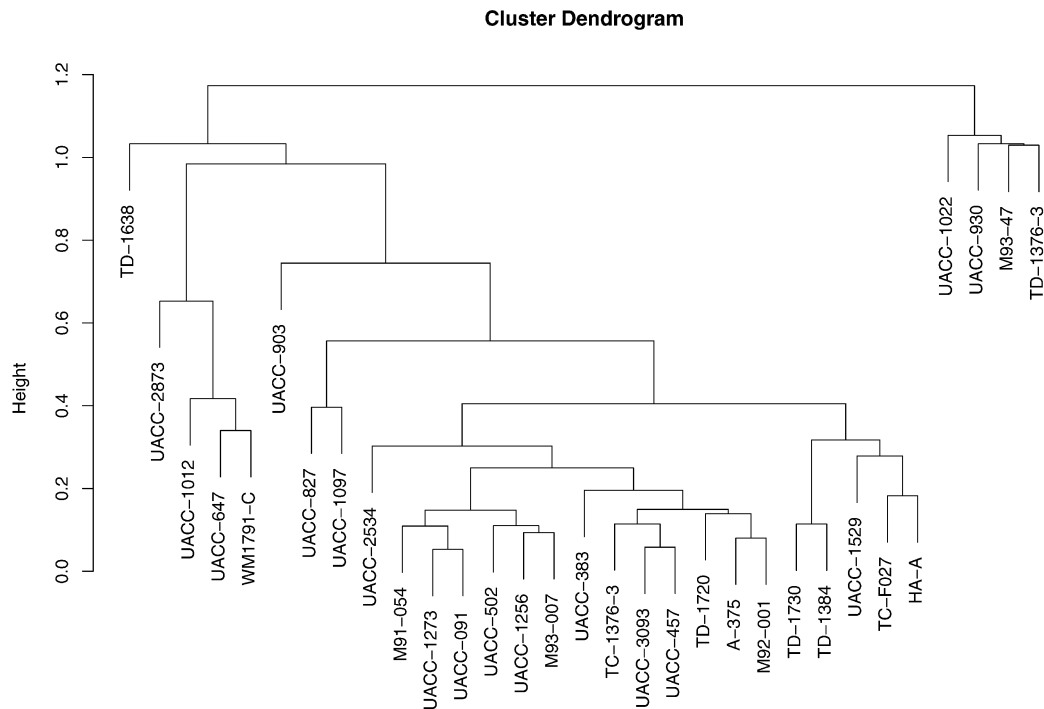


Fig. 1. HC dendrogram of gene expression data from Bittner *et al.* (2000). Average linkage clustering used.

The gene expression ratios are first transformed using principal components; the plot of the proportion of the total variance explained by each of the individual principal components relative to the total variance is given in Figure 2. Based on the graph, we chose to consider the first two principal components in the analysis. First, the agglomerative HC was performed. As was mentioned earlier, this step requires a specification of the covariance structure for each cluster. We present the results for the exchangeable model with unequal volume per cluster (i.e. $\Sigma_k = \lambda_k \mathbf{I}$ for $k = 1, \dots, K$). A plot of the clustering for two groups $K = 2$ based on agglomerative HC is given in Figure 3. Based on these graphs, we find that there is good separation between the two clusters. Similar results were obtained using other covariance models in the agglomerative HC algorithm (data not shown).

These results were then used as the input for maximum likelihood estimation of (1), as well as for computation of Bayes factors for model selection. A variety of cluster covariance structures were examined. We considered mixture models of the form (1) that had between one and six components. For larger numbers of components in the mixture model, we found that there were convergence problems with the EM algorithm. It becomes harder to estimate the variance parameters for models with larger numbers of components. This issue will be addressed in more detail in Section **Discussion**. In Table 2, we have

Table 2. Mixture model-based clustering results of samples for melanoma data

Variance model	Number of components	BIC
$\lambda \mathbf{I}$	1	-513.67
$\lambda_k \mathbf{I}$	1	-513.67
$\lambda \mathbf{DAD}^T$	2	-514.08
$\lambda \mathbf{D}_k \mathbf{A D}_k^T$	3	-514.02
$\lambda_k \mathbf{D}_k \mathbf{A D}_k^T$	2	-516.09
$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$	1	-518.45

listed the mixture model chosen using Bayes factors under a variety of covariance structures for (3). Based on the results of this table, the mixture model selected in most of the analyses was one in which the samples came from one population. This implies that there are no real clusters in the data. A second model chosen was one in which the samples came from two populations. A breakdown of the samples by cluster is given in Table 3.

In addition, we utilized the CAST algorithm (Ben-Dor *et al.*, 1999) for clustering the samples. This nonhierarchical algorithm explicitly makes no assumption on the number of clusters in the data; however, it requires the specification of a threshold parameter t . We set $t = 0.25$

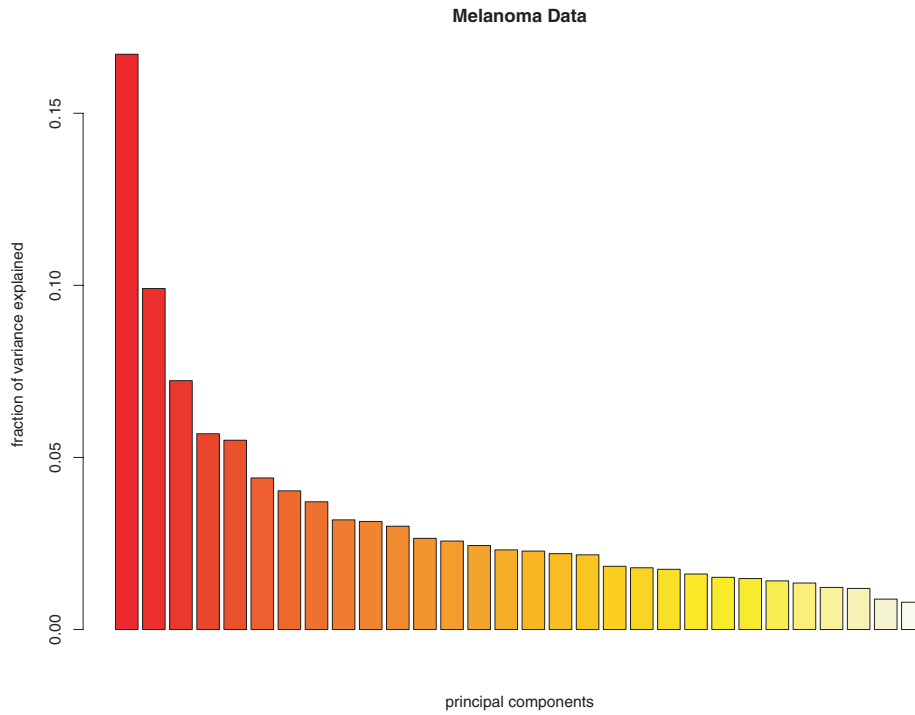


Fig. 2. Graph of proportion of variance explained by principal components for melanoma data.

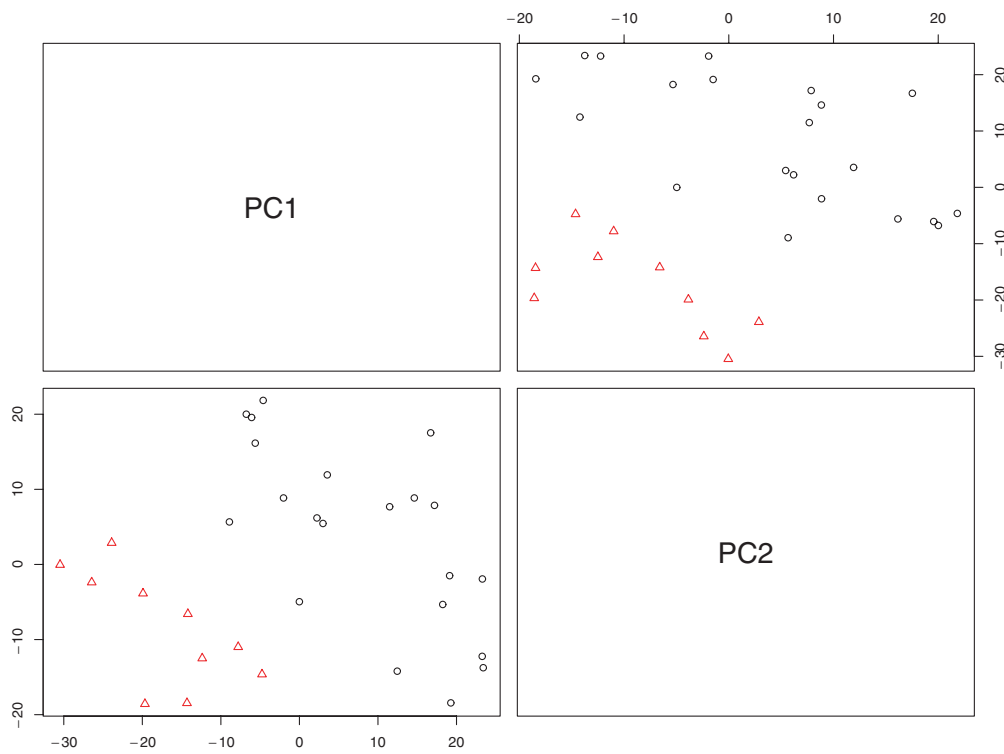


Fig. 3. Clustering of samples in melanoma data using model-based clustering; first two principal components used. Two groups are represented by circles and triangles.

Table 3. Melanoma clusters from mixture model-based clustering

Cluster 1	Cluster 2
UACC-2873	TC-F027
UACC-1012	HA-A
UACC-1529	TD-1720
UACC-647	TD-1638
WM1791-C	TD-1730
UACC-827	TD-1376-3
UACC-930	TC-1376-3
UACC-903	TD-1384
UACC-1097	UACC-1022
M93-47	A-375
	UACC-3093
	UACC-383
	UACC-457
	M92-001
	UACC-2534
	UACC-1273
	UACC-1256
	UACC-502
	UACC-091
	M91-054
	M93-007

and obtained three clusters in the data. Another mixture model-based approach for clustering samples is given by the AutoClass method (Cheeseman and Stutz, 1996). In their procedure, the number of clusters is determined through a combination of random starting values and a heuristic fitting method based on the log-normal distribution. For clustering samples, we need to reduce the dimension of the gene expression data; we utilize the singular value decomposition referred to earlier. The AutoClass approach yields six clusters.

Given the results of the cluster reliability methods and the mixture model-based clustering analyses, it suggests one of two possible conclusions. First, there may be no real distinct subtypes in the 31 melanoma samples. Second, the 19 melanoma cluster found by Bittner *et al.* (2000) should be broadened to include more melanoma samples, potentially as large as 21–23 melanoma samples.

Prostate cancer data

In this set of experiments, reported by our group (Dhanasekaran *et al.*, 2001), the goal was to utilize cDNA microarray technology in order to identify candidate cancer biomarkers or genes involved in prostate carcinogenesis. Two reference samples were used as the baseline sample for the microarray, a pool of normal adjacent prostate tissue and a pool of normal prostate tissue from patients without prostate pathology. For the analyses reported here, we consider samples profiled using the normal adjacent pool as the reference. There

are 26 samples: 5 benign prostate hyperplasia samples, 1 prostatitis sample, 3 normal adjacent prostate samples, 10 clinically localized prostate cancer samples and 7 metastatic prostate cancer samples.

We considered gene expression data from 3955 genes across the 26 samples. The original arrays developed contained 9984 elements, including 5000 known genes from the Research Genetics human cDNA clone set, 4400 ESTs and 500 control elements (which include genomic human, rat and yeast DNAs). The following preprocessing steps were taken to filter the number of genes from 9984 to 3955. First, genes where at least one channel did not give a signal at least 350 units above background were excluded. Next, genes that did not elicit a signal across at least 75% of the arrays were not included in the analysis. In addition, genes with absolute ratios less than one were excluded from the dataset. Finally, we only considered genes that had no missing measurements.

In order to normalize for differences in gene expression between tissue types, we fit the following gene-specific analysis of variance model for each of the 3955 genes:

$$\log(Y_{ij}) = \eta_j + \gamma_i + \epsilon_{ij},$$

where η_j denotes the baseline gene expression ratio for the j th gene, γ_i is the fixed effect corresponding to tissue type (metastatic, locally advanced prostate cancer and other), and ϵ_{ij} are iid mean zero error terms, $i = 1, \dots, 26$, $j = 1, \dots, 3955$. The residuals from the estimated models were used for the clustering analyses. The methods of Calinski and Harabasz (1974); Hartigan (1975) and Krzanowski and Lai (1985) estimate two clusters in the data.

First, clustering of the samples for this dataset is considered. We again apply principal components analysis to the gene expression measurements. In Figure 4, we show a plot of the proportion of the variance explained by the principal components. Based on this graph, we again selected three principal components for further analysis. A procedure similar to that for the melanoma was followed. First, agglomerative HC was performed. Figure 5 depicts the results of this method for two groups, where the covariance structure for each group was $\Sigma_k = \lambda_k \mathbf{DAD}^T$ $k = 1, 2, 3$. In contrast to the melanoma dataset, the groups are not as well-separated. The agglomerative HC results did not appear to be very sensitive to the covariance structure. The next stage involves maximum likelihood estimation using the EM algorithm and model selection using Bayes factors. Again, there were some convergence issues estimating variance parameters for mixture models with more than four components. In Table 4, we summarize the number of clusters that yielded the largest BIC value under several covariance structures for the clusters. Interestingly, there is more evidence for

Table 4. Mixture model-based clustering results of samples for prostate data

Variance model	Number of components	BIC
$\lambda \mathbf{I}$	1	-446.32
$\lambda_k \mathbf{I}$	2	-443.28
$\lambda \mathbf{DAD}^T$	6	-457.46
$\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$	5	-419.02
$\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$	4	-430.46
$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$	1	-399.06

multiple clusters with these data than for the melanoma data. However, the number of clusters depends on the particular covariance structure used. The CAST algorithm with a threshold of $t = 0.25$ yields four clusters in the data. Applying AutoClass to the principal components yields an estimated five clusters in the data.

One of the analyses performed in Dhanasekaran *et al.* (2001) involved determining functional groups of genes. We used the proposed mixture modelling method for genes to achieve this goal. We begin by applying a k -means clustering to the gene expression data to create a coarse subdivision of the data and to substantially increase the computing efficiency of the model-based clustering method. For this analysis, we took $k = 1000$. The partition from the k -means algorithm was then used as an input for the model-based agglomerative HC and subsequent steps of the model-based clustering. Because of convergence problems, we used the variance structure $\Sigma_k \equiv \lambda_k \mathbf{I}$ in the analyses. Approximate Bayes factors were used to determine the best-fitting mixture models. Based on this criterion, a fifteen component mixture model was chosen. We then looked at the clusters individually to determine if there was any functional similarity within each group.

One of the interesting findings was that hepsin, a transmembrane serine protease, was found to be clustered with several genes known to be implicated in various cancer pathways. A subset of these genes is given in Table 5. ‘Guilt by association’ would suggest a potential role for hepsin in prostate carcinogenesis. Subsequent follow-up experiments using high-density tissue microarrays demonstrated a correlation between hepsin protein expression with prostate cancer and Prostate Specific Antigen (PSA) failure. These analyses are described in further detail in Dhanasekaran *et al.* (2001).

We first examined the sensitivity of the results to the choice of K . In particular, we examined the concordance of the genes in the list in Table 5 for the other choices of K . We considered $K = 250, 1000$ and 2000 . We found that for $K = 250$, seven of the genes were grouped with hepsin; only the FYN oncogene was grouped in a separate cluster. For $K = 2000$, all eight genes were grouped with hepsin.

Table 5. Putative cancer-related genes from hepsin cluster

p53-induced protein
v-raf-1 murine leukemia viral oncogene homolog 1
Tumor necrosis factor, alpha-induced protein 6
Transforming growth factor, beta 1
Tumor necrosis factor (ligand) superfamily, member 10
v-jun avian sarcoma virus 17 oncogene homolog
FYN oncogene related to SRC, FGR, YES
Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog

Table 6. CPU time (in seconds) for mixture modelling of prostate cancer genes

Step	K		
	250	1000	2000
k -means	3.9	6.6	8.5
Agglomerative HC	57.3	74.6	95.0
EM-algorithm	800	1200	2800

We also investigated the computation time required for each of the steps in the mixture modelling methodology for these values of K . The CPU times are given for a Windows-based operating system on a laptop with a Pentium 450 processor. The times required for each of the steps in fitting the mixture model are given in Table 6. Because of the preprocessing using the k -means algorithm, we can increase the speed of the mixture modelling estimation procedure. If we wish to use a coarser preprocessing (e.g. $K = 250$), we can increase the computing efficiency of the estimation procedure.

Finally, we examined the sensitivity of the gene clustering with respect to the choice of initial preprocessing clustering algorithm. We used the Partitioning Around Medoids (PAM) algorithm (Kauffman and Rousseuw, 1990) as the preprocessing clustering algorithm with $K = 250, 1000$ and 2000 . We examined the concordance with the k -means-based procedure using the list of genes in Table 5. We found perfect concordance between the k -means and PAM procedures for each value of K .

DISCUSSION

Here we have developed a mixture modelling approach for the analysis of data from cDNA microarray experiments. There are aspects of these studies that pose unique challenges in the application of these methods; we have proposed new algorithms to address these issues.

These methods can serve as a complementary analysis to standard HC algorithms. An attractive feature of the mixture modelling approach is that a strength of evidence measure for the number of true clusters in

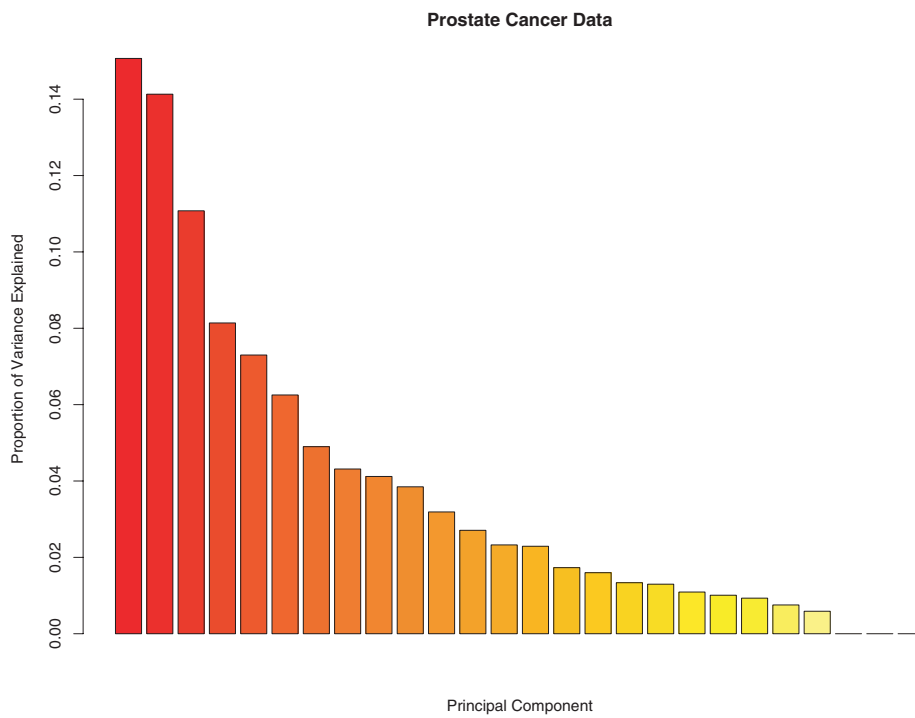


Fig. 4. Graph of proportion of variance explained by principal components for prostate cancer data.

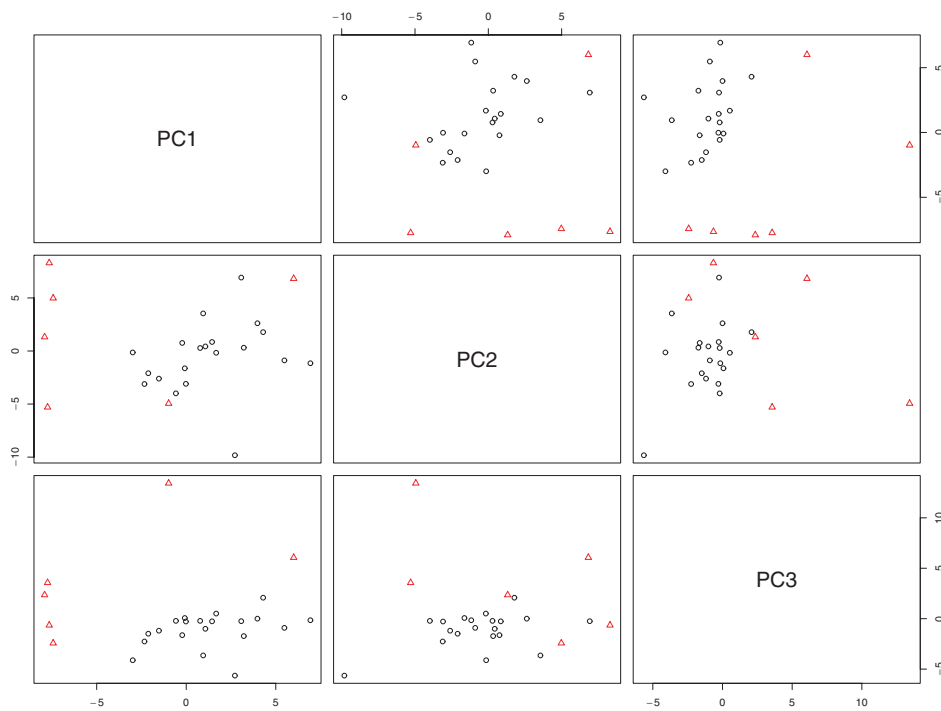


Fig. 5. Clustering of samples in prostate cancer data using model-based clustering; first three principal components used. Two groups are represented by circles and triangles.

the data is computed. This assessment of reliability of clustering output is often an important question to biologists considering data from microarray studies.

In the analyses of the two datasets, there were some convergence issues involved with the EM algorithm. This is because it is difficult to estimate variance parameters for clusters with a small number of samples. One way of addressing this problem would be to use a fully Bayesian estimation procedure. Incorporation of a prior distribution for the variance matrices would avoid the convergence difficulties.

We have applied the mixture model methodology to two molecular profiling studies in cancer (Bittner *et al.*, 2000; Dhanasekaran *et al.*, 2001). Another type of microarray experiment is the time-course study, in which the gene expression of the same population of cells is measured at a different number of time points. The methods proposed here would not be applicable to this setting; alternative clustering methods are needed instead.

A desirable feature of the mixture modelling approach presented here is that it is based on a statistical model. Another problem of interest is the classification of ESTs into known classes of genes whose function are known (e.g. proteases, adhesion proteins, transcription factors) using microarray data. An approach to accomplishing this task using support vector machines was described by Brown *et al.* (2000). While that method is algorithmic, it should be possible to extend the mixture modelling approach employed here to do this hybrid classification.

ACKNOWLEDGEMENTS

This work was supported in part by a pilot research grant from the University of Michigan Bioinformatics Program (to D.G. and A.M.C.) and a Career Development Award from the University of Michigan Prostate SPOR (to A.M.C.), National Cancer Institute.

REFERENCES

- Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
- Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.
- Banfield, J.D. and Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.
- Barash, Y. and Friedman, N. (2001) Context-specific Bayesian clustering for gene expression data. *Proceedings of the Fifth Annual International Conference on Computational Biology*, 22–25 April, Montreal, Quebec, Canada.
- Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.
- Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sempas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D. and Sondak, V. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.
- Boyles, R.A. (1983) On the convergence of the EM algorithm. *J. R. Stat. Soc. B*, **45**, 47–50.
- Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. Jr and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
- Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.*, **21**(Suppl.), 33–37.
- Calinski, R.B. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Commun. Stat.*, **3**, 1–27.
- Campbell, J.G., Fraley, C., Stanford, D., Murtagh, F. and Raftery, A.E. (1999) Model-based methods for real-time textile fault detection. *Int. J. Imaging Syst. Technol.*, **10**, 339–346.
- Cheeseman, P. and Stutz, J. (1996) Bayesian classification (Auto-Class); theory and results. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, pp. 61–83.
- Dasgupta, A. and Raftery, A.E. (1998) Detecting features in spatial point processes with clutter via model-based clustering. *J. Am. Stat. Assoc.*, **79**, 762–771.
- 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.
- Dhanasekaran, S., Barrette, T., Ghosh, D., Shah, R., Kurachi, K., Pienta, K., Rubin, M.A. and Chinnaiyan, A.M. (2001) Molecular profiling of prostate cancer: delineation of candidate biomarkers and regulatory genes, submitted.
- Eisen, M.B., Spellman, 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.
- Everitt, B.S. (1993) *Cluster Analysis*. Arnold, London.
- Fraley, C. and Raftery, A.E. (1999) MCLUST: software for model-based cluster analysis. *J. Classification*, **16**, 297–306.
- Goldstein, D., Ghosh, D. and Conlon, E. (2001) Statistical issues in the clustering of gene expression data. *Statistica Sinica*, in press.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Hartigan, J. (1975) *Clustering Algorithms*. Wiley, New York.
- Holmes, I. and Bruno, W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 19–23 August, La Jolla, California.
- Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
- Kauffman, L. and Rousseeuw, P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- Krzanowski, W.J. and Lai, Y.T. (1985) A criterion for determining the

- number of groups in a data set using sum of squares clustering. *Biometrics*, **44**, 23–34.
- Lipshutz,R.J., Fodor,S.P., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nature Genet.*, **21** (Suppl.), 20–24.
- Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotech.*, **14**, 1675–1680.
- MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium*, vol 1, pp. 281–297.
- Perou,C.M., Sorlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H. and Ak-slen,L.A. *et al.* (2000) Molecular portraits of human breast tumors. *Nature*, **406**, 747–752.
- Roeder,K. and Wasserman,L. (1997) Practical Bayesian density estimation using mixtures of normal. *J. Am. Stat. Assoc.*, **92**, 894–902.
- Schuchhardt,J., Beule,D., Malik,A., Wolski,E., Eickhoff,H., Lehrach,H. and Herzog,H. (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res.*, **28**, E47.
- Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.
- Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
- Wu,C.F. J. (1983) On the convergence of the EM algorithm. *Ann. Stat.*, **11**, 95–103.
- Yang,Y.H., Dudoit,S., Luu,P. and Speed,T.P. (2001) Normalization for cDNA microarray data. *Technical Report*. Department of Statistics, UC-Berkeley.