Contents

1 Clustering of Microarray Data via Mixture Models **1**

1.1 Introduction 1

1.2 Clustering of Microarray Data 5

1.3 Notation 6

1.4 Clustering of Tissue Samples 9

1.5 The EMMIX-GENE Clustering Procedure 10

1.6 Clustering of Gene Profiles 17

1.7 EMMIX-WIRE 19

1.8 ML Estimation via the EM algorithm 22

1.9 Model Selection 25

1.10 Example: Clustering of Time-Course Data 26

1.11 Discussion 29

1 Clustering of Microarray Data via Mixture Models

1.1 INTRODUCTION

The widespread use of DNA microarray technology (Eisen and Brown, 1999) to perform experiments on thousands of gene fragments in parallel has led to an explosion of expression data. A variety of multivariate analysis methods have been used to explore these data for relationships among the genes and the tissue samples. Cluster analysis has been one of the most frequently used methods for these purposes. It is an exploratory technique that attempts to find groups of observations that have similar values on a set of variables. Sometimes emphasis is placed on the distinction between the search for naturally occurring clusters and the division of the entities into a given number of groups, where there is no implication that the resulting groups are in any sense a natural division of the data; see, for example, Hand and Heard (2005). But often there is no emphasis, particularly as most methods for finding natural clusters are also useful for segmenting the data.

Agglomerative hierarchical clustering (encompassing single-, complete-, and average-linkage variants), k-means clustering, and self-organizing maps (SOM) have been the most widely used methods. Eisen et al. (1998) was the first to apply cluster analysis to microarray data, using an agglomerative hierarchical method using av-

erage linkage with a correlation-based metric, or equivalently, the Euclidean metric after standardization of the data.

More recently, increasing attention is being given to model-based methods of clustering of microarray data (Ghosh and Chinnaiyan, 2002; Yeung et al., 2001; McLachlan et al., 2002; Medvedovic and Sivaganesan, 2002), among others.

A useful way to think about the different clustering procedures is in terms of the shape of the clusters produced (Reilly et al., 2005). Many clustering methods assume that the appropriate distance function (metric) is known (for example, they may use Euclidean distance). But clearly, it would be more appropriate to use a metric that depends on the shape of the clusters. As pointed out by Coleman et al. (1999), 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. The majority of the existing clustering methods assume that a similarity measure or metric is known *a priori*; often the Euclidean metric is used. In particular, k-means effectively uses the Euclidean metric, as it can be viewed as being a "hard" version of the mixture clustering procedure based on a mixture in equal proportions of multivariate normal components with a common spherical covariance matrix. In the absence of any prior knowledge on the metric, it is reasonable to adopt a clustering procedure that is invariant under affine transformations of the data; that is, invariant under transformations of the data $\boldsymbol{y}$ of the form,

$$\boldsymbol{y} \rightarrow \boldsymbol{C}\boldsymbol{y} + \boldsymbol{a}, \tag{1.1}$$

where $\boldsymbol{C}$ is a nonsingular matrix. One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or t densities, is that the implied clustering is invariant under affine transformations of the data (that is, under operations relating to changes in location, scale, and rotation of the data). Thus the clustering process does not depend on irrelevant factors such as the units of measurement or the orientation of the clusters in space. If the clustering of a procedure is invariant under (1.1) for only diagonal $\boldsymbol{C}$, then it is invariant under change of measuring units but not rotations. But as commented upon by Hartigan (1975), this form of invariance is more compelling than affine invariance.

In this chapter, we shall focus on a model-based approach to the clustering of microarray data using mixtures of normal distributions, which are commonly used in statistics; see, for example, Ganesalingam and McLachlan (1978), McLachlan and Basford (1988), Banfield and Raftery (1993), Fraley and Raftery (1998, 2002), and McLachlan and Peel (2000). As noted by Aitkin et al. (1981), "Clustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory." Previously, Marriott (1974, 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." More recently, Yeung et al. (2001) noted that "in the absence of a well-grounded statistical model, it seems difficult to define what is meant by a 'good' clustering algorithm or the 'right' number of clusters."

The normal mixture model-based approach is to be applied here in a nonhierarchical manner, as there is no reason why the clusters of tissues or genes should be hierarchical in nature. It is true that if there is a clear, unequivocal grouping, with little or no overlap between the groups, any method will reach this grouping. But as pointed out by Marriott (1974), "hierarchical methods are not primarily adapted to finding groups." For instance, if the division into $g = 2$ groups given by some hierarchical method is optimum with respect to some criterion, then the subsequent division into $g = 3$ groups is unlikely to be so. This is due to the restriction that one of the groups must be the same in both the $g = 2$ and $g = 3$ clusterings. As explained by Marriott (1974), this restriction is not a natural one to impose if the purpose is to find a natural grouping of the data. As advocated by Marriott (1974, Page 67), "it is better to consider the clustering problem *ab initio*, without imposing any conditions."

Another attractive feature of the use of mixture models for clustering is that the question of the number of clusters can be can be formulated in terms of a criterion or a test for the smallest number of components in the mixture model compatible with the data. One such criterion is the Bayesian information criterion (BIC) of Schwarz (1978), while a test can be carried out on the basis of the likelihood ratio statistic λ.

One potential drawback with the normal mixture model-based approach to clustering 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.

In practice, consideration has to be given to the problem of relatively large local maxima that occur as a consequence of a fitted component having a very small (but nonzero) variance for univariate data or generalized variance (the determinant of the covariance matrix) for multivariate data. Such a component corresponds to a cluster containing a few data points either relatively close together or almost lying in a lower-dimensional subspace in the case of multivariate data. There is thus a need to monitor the relative size of the fitted mixing proportions and of the component variances for univariate observations, or of the generalized component variances for multivariate data, in an attempt to identify these spurious local maximizers. One situation where an apparent spurious solution would be of practical interest is where one (or more) of the fitted components correspond to a small number of points that are distant from the rest of the points.

1.2 CLUSTERING OF MICROARRAY DATA

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 other concerns the clustering of the genes on the basis of the tissues. This duality is quite common. 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.

In clustering microarray data, the clusters of tissues can play a useful role in the discovery and understanding of new subclasses of diseases. The clusters of genes

obtained can be used to search for genetic pathways or groups of genes that might be regulated together. Also, in the first problem, we may wish first to summarize the information in the very large number of genes by clustering them into groups (of hyperspherical shape), which can be represented by some metagenes, such as the group-sample means. We can then carry out the clustering of the tissues in terms of these metagenes. As noted by Pollard and van der Laan (2002), most research on these two problems has been carried out with them considered separately rather than simultaneously. They propose a statistical framework for two-way clustering; see also Getz et al. (2000) and the references therein for earlier approaches on this problem.

We firstly consider the clustering of tissue samples, using the EMMIX-GENE procedure of McLachlan et al. (2002). For the clustering of the gene profiles, we shall describe a mixture model with random effects, EMMIX-WIRE (**EM**-based **MIX**ture analysis **WI**th **R**andom **E**ffects), as developed recently by Ng et al. (2006a). More information about these programs can be found at the web addresses `http://www.maths.uq.edu.au/~gjm/emmix-gene/` and `http://www.maths.uq.edu.au/~gjm/emmix/emmix.html`.

1.3 NOTATION

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 the $N \times M$ matrix. For each tissue, we can consider the expression levels of the N genes, called its *expression signature*. Conversely, for each gene, we can consider its expression levels across the different tissue samples, called its *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 expression levels are taken to be the measured (absolute) intensities for Affymetrix oligonucleotide arrays, whereas for the spotted arrays (cDNA or oligonucleotide arrays), are taken to be the ratios of sample versus control intensities, represented by the Cy5-channel (red) and Cy3-channel (green) images (see, for example, Dudoit et al. 2002). It is assumed that one starts the clustering process with preprocessed (relative) intensities, such as those produced by RMA (for Affy data), loess-modified log ratios, or differences of logged/generalized-logged data; see, for example, Parmigiani et al. (2003), Huber et al. (2003), Irizarry et al. (2003), Rocke and Durbin (2003), and Speed (2003). The $N \times M$ matrix is portrayed in Figure 1.1, where each sample represents a separate microarray experiment and generates a set of N expression levels, one for each gene.

In the sequel, we shall use the vector $\boldsymbol{y}_j$ to represent the measurement (feature observation) on the jth entity to be clustered. In the context of the classification of the tissues on the basis of the gene expressions, we can represent the $N \times M$ matrix $\boldsymbol{A}$ of gene expressions as

$$\boldsymbol{A} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M), \tag{1.2}$$

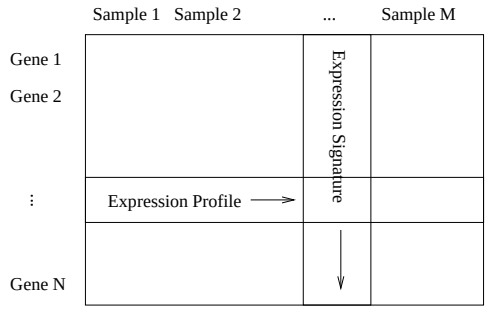

Fig. 1.1 Gene expression data from M microarray experiments represented as a matrix of expression levels with the N rows corresponding to the N genes and the M columns to the M tissue samples.

where the feature vector $\boldsymbol{y}_j$ (the *expression signature*) contains the expression levels on the N genes in the jth experiment $(j = 1, \ldots, M)$. The latter is a nonstandard problem in parametric cluster analysis because the dimension of the feature space (the number of genes) is typically much greater than the number of observations (the number of tissues).

In the context of the clustering of the genes on the basis of the tissues, we can represent the transpose of the matrix $\boldsymbol{A}$ in terms of the feature vectors as

$$\boldsymbol{A}^T = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N), \tag{1.3}$$

where the feature vector $\boldsymbol{y}_j$ (the *expression profile*) contains the expression levels on the M tissues on the jth gene $(j = 1, \ldots, N)$. For this clustering problem, the number of observations (the number of genes) is very large relative to the dimension of the feature space (the number of tissues), and so in this sense it falls in the standard

framework. However, it is not really a standard problem, as not all the genes are independently distributed.

1.4 CLUSTERING OF TISSUE SAMPLES

In the standard setting of a model-based cluster analysis, the n observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ to be clustered are taken to be independent realizations where the sample size n is much larger than the dimension p of each vector $\boldsymbol{y}_j$,

$$n >> p. \tag{1.4}$$

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

In the cluster analysis of the M tissue samples on the basis of the N genes, we have $n = M$ and $p = N$. Thus the sample size n will be typically small relative to the dimension p, causing estimation problems under the normal mixture model,

$$f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \phi(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1.5}$$

where $\phi(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p-dimensional normal density function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Psi}$ is the vector of unknown parameters. This is because the g-component normal mixture model (1.5) with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$.

An obvious way to handle the very large number of genes is to perform a principal component analysis and carry out the cluster analysis on the basis of the leading components. The shortcomings of a PCA in such a context is that the leading components need not necessarily reflect the direction in the feature space best for revealing the group structure of the tissues. This is because it is concerned with the direction of maximum variance, which is composed of variance within the clusters and variance between the clusters. If the latter are relatively large, then the leading components may not be so useful for the purposes of cluster analysis. But with the analysis of microarray data, this problem is compounded by the very large number of genes and their associated noise. Thus artificial directions can result from noisy genes and highly correlated ones. Consequently, a potential problem with a PCA is the determination of an appropriate number of principal components (PCs) useful for clustering. A common practice is to choose the first few leading components. But it may not be clear where to stop and whether some of these components are caused by some artifact or noises in the data. An excellent account of these problems may be found in Liu et al. (2003). They have developed a Bayesian approach to model-based clustering which after an initial PCA simultaneously clusters the observations and selects "informative" variables or components for the cluster analysis.

1.5 THE EMMIX-GENE CLUSTERING PROCEDURE

The EMMIX-GENE procedure handles the problem of a high-dimensional feature vector by using mixtures of factor analyzers whereby the component correlations

between the genes are explained by their conditional linear dependence on a small number q of latent or unobservable variables specific to each component. In practice we may wish to work with a subset of the available genes, particularly as the fitting of a mixture of factor analyzers will involve a considerable amount of computation time for an extremely large number of genes. Indeed, the simultaneous use of too many genes in the cluster analysis may serve only to create noise that masks the effect of a smaller number of genes. 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; see the recent papers of Pollard and van der Laan (2002) and Friedman and Meulman (2004) on this point. As explained in Belitskaya-Levy (2006), 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 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 in terms of the likelihood ratio test statistic. The second step clusters the retained genes N_o into groups on the basis of Euclidean distance so that highly correlated

genes are clustered into the same group. The third and final step of the EMMIX-GENE procedure considers the clustering of the tissues by fitting mixtures of normal distributions or factor analyzers. It can be implemented either by considering the groups of genes simultaneously on the basis of their means or by considering the groups individually on the basis of all or a subset of the genes in a given group. We now describe these three steps in more detail.

1.5.1 Step 1: Screening of Genes

In step 1 of EMMIX-GENE, we screen the genes by attempting to delete those genes that individually are of little use in clustering the tissue samples into two groups. This screening is undertaken in the absence of tissue samples that are of known classification. The relevance of a gene for clustering the tissue samples can be assessed on the basis of the value of $-2\log\lambda$, where λ is the likelihood ratio statistic for testing $g = 1$ versus $g = 2$ components in the mixture model. In order to reduce the effect of atypically large observations on the value of λ, we fit mixtures of t components with their degrees of freedom inferred from the data. However, the use of t components in place of normal components still does not eliminate the effect of outliers on inference of the number of groups in the tissue samples. For example, suppose that for a given gene there is no genuine grouping in the tissues, but that there are a small number of gross outliers. Then a significantly large value of λ might be obtained, with one component representing the main body of the data (and providing robust estimates of their underlying distribution) and the other representing

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

In light of the above, the EMMIX-GENE software automatically assesses the relevance of each of the N genes by fitting one- and two-component t mixture models to the expression data over the M tissues for each gene considered individually. If $-2\log\lambda$ is greater than a specified threshold b_1,

$$-2\log\lambda > b_1 \tag{1.6}$$

then the gene is taken to be relevant provided that

$$s_{\min} \geq b_2, \tag{1.7}$$

where $s_{\min}$ is the minimum size of the two clusters implied by the two-component t mixture model and b_2 is a specified threshold. If (1.6) holds but (1.7) does not for a given gene, then the three-component t mixture model is fitted to the tissue samples on this gene, and the value of $-2\log\lambda$ calculated for the test of $g = 2$ versus $g = 3$. If (1.6) holds for this value of $-2\log\lambda$, the gene is selected as being relevant (provided at least two of the three clusters implied by the $g = 3$ solution have sizes not less than b_2). Although the null distribution of $-2\log\lambda$ for $g = 2$ versus $g = 3$ is not the same as for $g = 1$ versus $g = 2$ components, it would appear to be reasonable here to use the same threshold (1.6). The null distribution of $-2\log\lambda$ for the test of the null hypothesis $H_0 : g = g_0$ versus the alternative hypothesis $H_1 : g = g_1$ is unknown (for finite sample sizes) for normal or t components (McLachlan and Peel,

2000, Chapter 6). In our applications of EMMIX-GENE, we have taken

$$b_1 = b_2 = 8. \tag{1.8}$$

The majority of genes in microarray data sets tend to exhibit near-constant expressions across samples (Dudoit and Fridlyand, 2002), and so many methods preselect genes by eliminating those with small variance. For example, the gene shaving methodology of Hastie et al. (2000) is concerned with the identification of small, homogeneous subsets of genes that have maximal variance across the tissue samples. As noted by Pollard and van der Laan (2002), genes with low variance can be equally interesting biologically, and so their two-way clustering procedure using hierarchical PAM (partitioning around medoids) is aimed at identifying clusters of genes with both low and high variance across tissues. The gene-selection procedure in EMMIX-GENE aims to identify genes whose distributions are not consistent with a single normal distribution, and so it can identify potentially valuable genes for clustering that can have both small and high variances across the tissues.

1.5.2 Step 2: Clustering of Genes: Formation of Metagenes

Concerning the end problem of clustering the tissue samples on the basis of the genes considered simultaneously, we could examine the univariate clusterings provided by each of the selected genes taken individually. But this would be rather tedious when a large number of genes have been selected. Thus with the EMMIX-GENE approach, there is a second (optional) stage for clustering the genes into a user-specified number (N_o) of groups by fitting a mixture in equal proportions of $g = N_o$

normal distributions with covariance matrices restricted to being equal to a multiple of the $(M \times M)$ identity matrix. That is, if the mixing proportions were fixed at 0.5, then it would be equivalent to using a soft version of k-means and grouping the genes in terms of the Euclidean distance between them. Since the gene-profiles have been normalized, they lie on the surface of the unit hypersphere. Thus, after each M-step of the EM algorithm, we normalize the updated estimates of the component means so that they lie on the surface of the unit hypersphere. More precisely, we could fit mixtures of von Mises-Fisher distributions as in Banerjee et al. (2006).

Each group (cluster) of genes can be represented by one or more M-dimensional profile vectors over the M tissues. We follow Huang (2003) in referring to these cluster representatives as *metagenes*. In EMMIX-GENE, we take the sample mean of the genes within a cluster to be the metagene representing the cluster. This strategy of using a linear combination of the genes within a cluster to represent it and so thereby reducing the dimension of the feature (gene) space also helps smooth out gene-specific noise through the aggregation within a cluster.

The groups of genes are ranked in terms of the likelihood ratio statistic calculated on the basis of the fitted mean of a group over the tissues for the test of a single versus two t components. This is provided that the minimum cluster size is greater than a specified threshold. Otherwise, such a group of genes would be put at the end of the list.

A heat map of genes in a group versus the tissues is provided for each of the groups where, in each group, the tissues can be left in their original order or rearranged

according to their cluster membership obtained by fitting a univariate t mixture model on the basis of the group mean. Alternatively, one could cluster the tissues by fitting a two-component mixture of factor analyzers on the basis of the genes within the group. Concerning the use of heat maps, they present a grid of colored points where each color represents a gene expression value for a gene in the tissue sample. They are used here primarily to exhibit similarities between groups or clusters of the tissue samples. Thus they are most effective in this role when the tissue samples have been grouped according to their group (cluster) memberships. Of course the heat maps are also useful in revealing similarities between the genes.

1.5.3 Step 3: Clustering of Tissues

If a clustering is sought on the basis of the totality of the genes, then it can be obtained by fitting a mixture model to these group means. However, it may be that the number of group means N_o is too large to fit a normal mixture model with unrestricted component-covariance matrices. In this circumstance EMMIX-GENE has the option on the third step that allows for the fitting of mixtures of factor analyzers. The use of mixtures of factor analyzers 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 q ($q < N$) unobservable factors. In addition to clustering the tissues on the basis of all of the genes, there may be interest in seeing if the different groups of genes lead to different clusterings of the tissues when each is

considered separately. For example, a subset of the genes may be all that is required to identify certain subtypes of the cancer being studied.

It can be seen from above that with the EMMIX-GENE procedure, the genes are being treated anonymously. That is, we do not incorporate existing biological information on the function of genes into the selection procedure. Spang (2003) infuses some biological context into an otherwise unsupervised learning task. He structures the feature space by using a functional grid provided by the Gene Ontology annotations.

1.6 CLUSTERING OF GENE PROFILES

In the remainder of this chapter, we consider the clustering of gene profiles with or without replication across some experimental conditions of interest. For this clustering problem, the number of observations n to be clustered is the number of genes $(n = N)$, which will usually be very large relative to the dimension p of the feature space $(p = M)$. In this sense it falls in the standard framework. However, this clustering problem is not straightforward as the profiles of the genes are not all independently distributed and the expression levels may have been obtained from an experimental design involving replicated arrays. Thus the standard normal mixture model (1.5) cannot directly be applied to cluster the gene profiles. This is because in unmodified form, this approach does not incorporate experimental design information such as disease status of the tissue samples in which the genes are measured in cross-sectional studies, covariate information such as the time ordering

of the gene measurements in time-course studies, or the structure of the replicated data as in longitudinal studies. Recently, Pan (2006) has proposed to incorporate known gene functions as prior probabilities in model-based clustering. But there is a need to develop further clustering procedures that are applicable to data from a wide variety of experimental designs. For example, microarray experiments are now being carried out with replication for capturing either biological or technical variability in expression levels to improve the quality of inferences made from experimental studies (Lee et al., 2000 and Pavlidis et al., 2003). Replicated measurements from each tissue sample (subject) are often interdependent and tend to be more alike in characteristics than data chosen at random from the population as a whole. Similarly, in time-course studies (Storey et al., 2005), where expression levels are measured under various conditions or at different time points, gene expressions obtained from the same condition (subject) are correlated.

Ng et al. (2006a) have developed a random-effects model that provides a unified approach to the clustering of genes with correlated expression levels measured in a wide variety of experimental situations. Their model is an extension of the normal mixture model (1.5) to account for the correlations between the gene profiles and to enable covariate information to be incorporated into the clustering process. Hence the model is applicable to longitudinal studies with or without replication, for example, time-course experiments by using time as a covariate, and to cross-sectional experiments by using categorical covariates to represent the different experimental classes. Ng et al. (2006a) have shown that their random-effects model EMMIX-WIRE (**EM**-based **MIX**ture analysis **WI**th **R**andom **E**ffects) can be fitted by maximum likelihood

via the Expectation–Maximization (EM) algorithm for which the E- and M-steps can be implemented in closed form. Hence their model can be fitted deterministically without the need for time-consuming Monte Carlo approximations.

In related work, Ng et al. (2006b) have applied this method of clustering to two real time-course datsets from the budding yeast (Saccharomyces cerevisiae) genome. They showed that the proposed method provided clusters of cell-cycle regulated genes that are supported by existing gene function annotations, and hence enables inference on regulatory interactions for the genetic network. Their approach was to search for regulatory control elements (activators and inhibitors) shared by the clusters of coexpressed genes, based on time-lagged correlations.

As noted by Bryan (2004) with the clustering of gene profiles, any clustering structure found may not be directly reflective of biological realities, but might be more due to the preprocessing of the data, which can create sparsely populated areas in the profile space as an artifact. In such situations, the clustering may still be of interest from the point of view of which genes are put together in the same cluster for various choices of the number of clusters.

1.7 EMMIX-WIRE

The EMMIX-WIRE procedure of Ng et al. (2006a) formulates a (multilevel) linear mixed-effects model (LMM) for the mixture components in which covariate information can be incorporated. It can be used for the clustering of correlated genes, based on expression microarray data obtained from various experimental designs

such as repeated measurement data and time-course data. Their proposed general random-effects model is formulated by incorporating both "gene" effects and "tissue" effects in the mixture modeling of the microarray data. This is in contrast to the mixed-effects models approaches in Celeux et al. (2005), Luan and Li (2003), and McLachlan et al. (2004) that involve only gene-specific random effects. Their methods thus require the independence assumption for the genes which, however, will not hold in practice for all pairs of genes (McLachlan et al., 2004).

With the EMMIX-WIRE procedure, it is assumed that the observed M-dimensional vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ are assumed to have come from a mixture of a finite number, say g, of components in some unknown proportions $\pi_1, \ldots, \pi_g$, which sum to one. Conditional on its membership of the hth component of the mixture, the vector $\boldsymbol{y}_j$ for the jth gene follows the model

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta}_h + \boldsymbol{U}\boldsymbol{b}_{hj} + \boldsymbol{V}\boldsymbol{c}_h + \boldsymbol{\epsilon}_{hj}, \tag{1.9}$$

where the elements of $\boldsymbol{\beta}_h$ (an M-dimensional vector) are fixed effects (unknown constants) modeling the conditional mean of $\boldsymbol{y}_j$ in the hth component, $\boldsymbol{b}_{hj}$ (a q_b-dimensional vector) and $\boldsymbol{c}_h$ (a q_c-dimensional vector) represent the unobservable gene- and cluster-specific random effects, respectively. The random effects $\boldsymbol{b}_{hj}$ and $\boldsymbol{c}_h$, and the measurement error vector $\boldsymbol{\epsilon}_{hj}$ are assumed to be mutually independent. In (1.9), $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are known design matrices of the corresponding fixed or random effects. The specification of (1.9) covers many general random-effects models for the clustering of correlated gene expression data arising from various microarray experiments, including those with replications. For example, let t be the number of

distinct tissues in the experiment. We are given for the jth gene a feature vector $\boldsymbol{y}_j = (\boldsymbol{y}_{1j}^T, \ldots, \boldsymbol{y}_{tj}^T)^T$, where $\boldsymbol{y}_{lj} = (y_{l1j}, \ldots, y_{lrj})^T$ contains the r replications on the jth gene from the lth tissue $(l = 1, \ldots, t)$. With respect to (1.9), $\boldsymbol{\beta}_h$ is a M-dimensional vector $(M = t)$ modeling the conditional mean of $\boldsymbol{y}_j$ in the hth component. Moreover, conditional on membership of the hth component, it is assumed that the random effects are shared among the repeated measurements of expression on the same gene from the same tissue ($\boldsymbol{b}_{hj}$ in (1.9) with $q_b = t$), along with the random effects that are shared among gene expressions from the same tissue ($\boldsymbol{c}_h$ in (1.9) with $q_c = M = tr$). The component-specific effects $\boldsymbol{c}_h$ for the tissues induce dependency among the gene-expression levels of genes from the same component and from the same tissue (correlated genes). By allowing the expression levels of the genes in a cluster to have their own and cluster-specific random-effects terms, there can be greater individual and collective variation, respectively, exhibited by the genes in the same cluster than otherwise possible under a fixed-effects model without gene- and cluster-specific random effects.

With the LMM, the distributions of $\boldsymbol{b}_{hj}$ and $\boldsymbol{c}_h$ are taken to be multivariate normal, $N_{q_b}(\mathbf{0}, \theta_{\mathbf{bh}} \boldsymbol{I}_{\mathbf{q_b}})$ and $N_{q_c}(\mathbf{0}, \theta_{\mathbf{ch}} \boldsymbol{I}_{\mathbf{q_c}})$, respectively, where $\boldsymbol{I}_{q_b}$ and $\boldsymbol{I}_{q_c}$ are identity matrices with dimensions being specified by the subscripts. The measurement error vector $\boldsymbol{\epsilon}_{hj}$ is also taken to be multivariate normal $N_M(\mathbf{0}, \boldsymbol{D}_{\mathbf{h}})$, where $\boldsymbol{D}_h = \text{diag}(\boldsymbol{W}\boldsymbol{\phi}_h)$ is a diagonal matrix constructed from the vector $(\boldsymbol{W}\boldsymbol{\phi}_h)$ with $\boldsymbol{\phi}_h = (\sigma_{h1}^2, \ldots, \sigma_{hq_e}^2)^T$ and $\boldsymbol{W}$ a known $M \times q_e$ zero-one design matrix. That is, we allow the hth component-variance to be different among the M microarray experiments.

1.8 ML ESTIMATION VIA THE EM ALGORITHM

We let $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1^T, \ldots, \boldsymbol{\psi}_g^T, \pi_1, \ldots, \pi_{g-1})^T$ be the vector of all the unknown parameters, where $\boldsymbol{\psi}_h$ is the vector containing the unknown parameters $\boldsymbol{\beta}_h$, θ_{bh}, θ_{ch}, and $\boldsymbol{\phi}_h$ of the hth component density $(h = 1, \ldots, g)$. Ng et al. (2006a) showed that the estimation of $\boldsymbol{\Psi}$ can be obtained by maximum likelihood (ML) via the EM algorithm of Dempster et al. (1977). The implementation of the E-step is straightforward for mixture models provided that the data can be treated as being independently distributed. In their model (1.9), the gene-profile vectors $\boldsymbol{y}_j$ are not all independently distributed as genes within the same cluster (that is, from the same component in the mixture model) and are allowed to be dependent due to the presence of the random-effects term $\boldsymbol{c}_h$ for the hth component in (1.9). However, this problem can be circumvented by proceeding conditionally on the random-cluster effects $\boldsymbol{c}_h$, as given these terms, the gene profile vectors $\boldsymbol{y}_j$ are all conditionally independent. In this way, Ng et al. (2006a) showed that the E- and M-steps can be carried out in closed form. In particular, we do not have to approximate the E-step by carrying out time-consuming Monte Carlo approximations.

Within the EM framework, each $\boldsymbol{y}_j$ is conceptualized to have arisen from one of the g components. We let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ denote the unobservable component-indicator vectors, where the hth element z_{hj} of $\boldsymbol{z}_j$ is taken to be one or zero according as $\boldsymbol{y}_j$ does or does not come from the hth component given $\boldsymbol{c}$, where $\boldsymbol{c} = (\boldsymbol{c}_1^T, \ldots, \boldsymbol{c}_g^T)^T$. We let $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T)^T$ denote the observed data and, correspondingly, put $\boldsymbol{z}^T = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_N^T)$. The ML estimation of the normal mixture of LMMs via the

EM algorithm can be formulated by treating the unobservable component-indicator variables $\boldsymbol{z}$ and the random effects $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_g^T)^T$ and $\boldsymbol{c}$ as missing data in the EM framework (Ng et al., 2004) where $\boldsymbol{b}_h = (\boldsymbol{b}_{h1}^T, \ldots, \boldsymbol{b}_{hN}^T)^T$ for $h = 1, \ldots, g$. Let $\boldsymbol{\epsilon}_h = (\boldsymbol{\epsilon}_{h1}^T, \ldots, \boldsymbol{\epsilon}_{hn}^T)^T$ for $h = 1, \ldots, g$. With

$$(\boldsymbol{y}^T, \boldsymbol{z}^T, \boldsymbol{b}^T, \boldsymbol{c}^T)^T$$

taken to be the complete data, it follows that the complete-data log likelihood is given, apart from an additive constant, by

$$\begin{aligned} \log L_c(\boldsymbol{\Psi}) \quad &= \quad \sum_{h=1}^{g} \Bigg[\sum_{j=1}^{n} z_{hj} \log \pi_h - \tfrac{1}{2} \Bigg\{ \sum_{j=1}^{n} z_{hj} q_b \log \theta_{bh} + \\ q_c \quad \log \theta_{ch} \quad &+ \sum_{j=1}^{n} z_{hj} \log |\boldsymbol{A}_h| \; + \frac{\boldsymbol{b}_h^T \boldsymbol{b}_h}{\theta_{bh}} + \frac{\boldsymbol{c}_h^T \boldsymbol{c}_h}{\theta_{ch}} + \boldsymbol{\epsilon}_h^T \boldsymbol{\Omega}_h \boldsymbol{\epsilon}_h \Bigg\} \Bigg], \end{aligned} \tag{1.10}$$

where

$$\boldsymbol{b}_h^T \boldsymbol{b}_h = \sum_{j=1}^{n} z_{hj} \boldsymbol{b}_{hj}^T \boldsymbol{b}_{hj}$$

and

$$\boldsymbol{\Omega}_h = \boldsymbol{I}_n \otimes \boldsymbol{A}_h^{-1}$$

for $h = 1, \ldots, g$, and hence

$$\boldsymbol{\epsilon}_h^T \boldsymbol{\Omega}_h \boldsymbol{\epsilon}_h = \sum_{j=1}^{n} z_{hj} \boldsymbol{\epsilon}_{hj}^T \boldsymbol{A}_h^{-1} \boldsymbol{\epsilon}_{hj}.$$

In the above, the sign $\otimes$ denotes the Kronecker product of two matrices. By consideration of (1.10), Ng et al. (2006a) showed that the E- and M-steps can be implemented in closed form.

To effect a probabilistic or an outright clustering of the genes into g components, we can condition on the cluster random-effects vector $\boldsymbol{c}_h$. As the latter is unobservable, we use its estimated conditional expectation given the observed data,

$$\hat{\boldsymbol{c}}_h = E_{\hat{\boldsymbol{\Psi}}}(\boldsymbol{c}_h \mid \boldsymbol{y}), \tag{1.11}$$

where $E_{\hat{\boldsymbol{\Psi}}}$ denotes taking expectation using the ML estimate $\hat{\boldsymbol{\Psi}}$ for the vector $\boldsymbol{\Psi}$ of unknown parameters. Since the genes within a cluster are independently distributed given $\boldsymbol{c}_h$, it suffices to effect a clustering with each gene considered individually in terms of its estimated posterior probabilities of component membership given its profile vector and $\boldsymbol{c}_h$, for $h = 1, \ldots, g$ and $j = 1, \ldots, n$. Using Bayes' theorem, the posterior probability that the jth gene belongs to the hth component given $\boldsymbol{y}_j$ and $\boldsymbol{c}$, $\tau(\boldsymbol{y}_j, \boldsymbol{c}; \boldsymbol{\Psi})$ can be expressed as

$$\begin{aligned} \tau(\boldsymbol{y}_j, \boldsymbol{c}; \boldsymbol{\Psi}) &= \mathrm{pr}\{Z_{hj} = 1 \mid \boldsymbol{y}_j, \boldsymbol{c}\} \\ &= \frac{\pi_h f(\boldsymbol{y}_j \mid z_{hj} = 1, \boldsymbol{c}_h; \boldsymbol{\psi}_h)}{\sum_{i=1}^{g} \pi_i f(\boldsymbol{y}_j \mid z_{ij} = 1, \boldsymbol{c}_i; \boldsymbol{\psi}_i),} \end{aligned} \tag{1.12}$$

where $f(\boldsymbol{y}_j \mid z_{hj} = 1, \boldsymbol{c}_h; \boldsymbol{\psi}_h)$ denotes the hth component density of $\boldsymbol{y}_j$ given the random effect $\boldsymbol{c}_h$. The log of this density is given by

$$\begin{aligned} \log f(\boldsymbol{y}_j \mid z_{hj} = 1, \boldsymbol{c}_h; \boldsymbol{\psi}_h) = -\tfrac{1}{2}\Big\{ &\log |\boldsymbol{B}_h| + \\ &(\boldsymbol{y}_j - \boldsymbol{X}\boldsymbol{\beta}_h - \boldsymbol{V}\boldsymbol{c}_h)^T \boldsymbol{B}_h^{-1} (\boldsymbol{y}_j - \boldsymbol{X}\boldsymbol{\beta}_h - \boldsymbol{V}\boldsymbol{c}_h) \Big\}, \end{aligned}$$

apart from an additive constant, is the log of the hth component density of $\boldsymbol{y}_j$ conditional on $\boldsymbol{c}_h$, where $\boldsymbol{B}_h = \boldsymbol{A}_h + \theta_{bh}\boldsymbol{U}\boldsymbol{U}^T$.

1.9 MODEL SELECTION

The specification of the random-effects components in the model (1.9) needs careful consideration. An identifiability problem could arise if the random-effects model is specified so that the design matrix $\boldsymbol{V}$ for the random effects $\boldsymbol{c}_h$ is the same as the $\boldsymbol{X}$ for the fixed effects $\boldsymbol{\beta}_h$. In their study, Ng et al. (2006a) were concerned with situations where the emphasis is on the grouping of the genes rather than on the number of clusters and their link with externally existing groups. That is, they were concerned primarily in finding which genes are put together in the same cluster for plausible choices of the number of components g in the mixture model. A guide to plausible values of g can be obtained using BIC (the Bayesian information criterion) of Schwarz (1978), whereby the number g of components in the mixture model is taken to minimize $-2\log L(\hat{\boldsymbol{\Psi}}) + d\log n$, and d denotes the number of parameters in the model. In the EM framework, $L(\boldsymbol{\Psi})$ is the incomplete-data likelihood function for $\boldsymbol{\Psi}$. However, as the gene-profile vectors $\boldsymbol{y}_j$ are not all independently distributed, this likelihood function $L(\boldsymbol{\Psi})$ is unable to be calculated directly by taking the product of the (marginal) densities of the $\boldsymbol{y}_j$. Ng et al. (2006a) suggested that $L(\boldsymbol{\Psi})$ be approximated by forming it as if all the $\boldsymbol{y}_j$ were independent. Another approach would be to use resampling methods (Efron and Tibshirani, 1993; McLachlan, 1987; McLachlan and Khan, 2004).

1.10 EXAMPLE: CLUSTERING OF TIME-COURSE DATA

To illustrate the EMMIX-WIRE approach to the clustering of gene profiles, Ng et al. (2006a) applied it to three representative data sets, each arising from different kinds of microarray experiments: time course data as in the yeast cell-cycle study of Spellman et al. (1988), data with repeated measurements as in the yeast galactose study of Ideker et al. (2000), and finally cross-sectional data involving two groups of tissues (tumuor and normal) as in the study of human colorectal carcinomas of Muro et al. (2003).

We report here their first example. By analyzing cDNA microarrays from yeast cultures synchronized by three independent methods over approximately two cell-cycle periods, Spellman et al. (1998) identified 800 yeast genes that meet an objective minimum criterion for cell cycle regulation. In their study, Ng et al. (2006a) considered the 18 α-factor (pheromone) synchronization where the yeast cells were sampled at 7 minute intervals for 119 minutes. They worked with a subset of 612 genes that had no missing expression data across any of the 18 time points. Their aim was to cluster the cell cycle-regulated genes based on the microarray expression data matrix of $N = 612$ rows (genes) and $M = 18$ columns (time points). They then analyzed the clusters so formed for common regulatory elements, as described by Spellman et al. (1998). With reference to (1.9), they took the design matrix $\boldsymbol{X}$ to be an 18×2 matrix with the $(l+1)$th row $(l = 0, \ldots, 17)$

$$(\cos(2\pi(7l)/\omega + \Phi) \quad \sin(2\pi(7l)/\omega + \Phi)),$$

where ω is the period of the cell cycle and Φ is the phase offset. They adopted here the least squares estimation approach considered by Booth et al. (2004) to obtain the cell cycle period $\omega = 53$ and the initial phase $\Phi = 0$ from the data set. For the design matrices of the random effects parts, they took $\boldsymbol{U} = \mathbf{1}_{18}$ and $\boldsymbol{V} = \boldsymbol{I}_{18}$. That is, it is assumed that there exist random gene effects b_{hj} with $q_b = 1$ and random temporal effects $(c_{h1}, \ldots, c_{hq_c})$ with $q_c = m = 18$. The latter introduce interdependency among expression levels within the same cluster obtained from the same time point. Also, they took $\boldsymbol{W} = \mathbf{1}_{18}$ and $\boldsymbol{\phi}_h = \sigma_h^2$ ($q_e = 1$) so that the component variances were common among the $m = 18$ experiments. The mixture model of LMMs was fitted to the data with $g = 4$ to $g = 15$ components. The number of components g was determined using BIC for model selection. It indicated here that there are twelve clusters.

The clustering results for $g = 12$ as obtained by Ng et al. (2006a) are given in Figure 1.2, where the expression profiles for genes in each cluster are presented. From Figure 1.2, it can be seen that the genes have very similar expression patterns within each cluster, except in clusters 4 and 7, where there is greater individual variation in some of the genes. This clustering result is different from Spellman's clustering, which was based on time of peak expression only.

For Clusters 1, 3, 10, 11, and 12 that show clear periodic expression patterns, Ng et al. (2006a) searched through the 700-bp upstream region of the start codon of each gene for the presence of binding site sequences for any known yeast cell cycle transcription factors like MBF, SBF, Mcm1p-containing factors, and Swi5p factors.

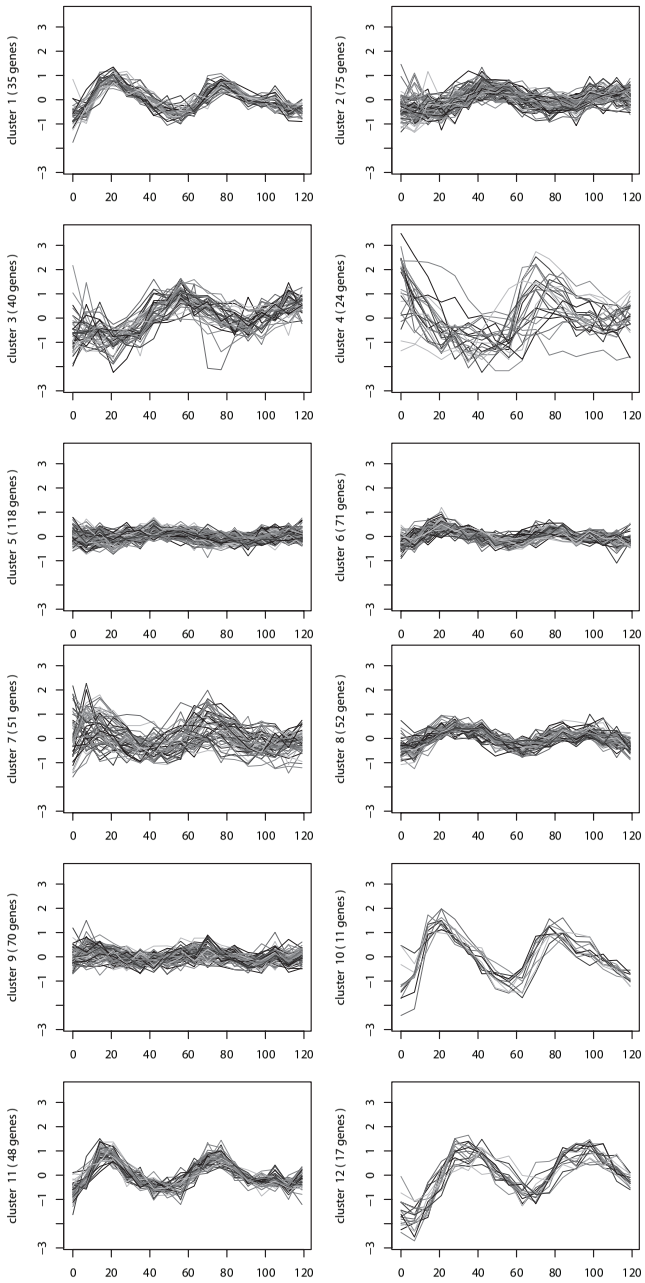

Fig. 1.2 Clustering results for the yeast cell-cycle data. For all the plots, the x-axis is the time point and the y-axis is the gene-expression level.

Cluster	No. of genes	Binding site	Regulator	Peak expression
1	35	ACGCGT	MBF, SBF	G1
3	40	MCM1 + SFF	Mcm1p + SFF	G2/M
10	11	ACGCGT	MBF, SBF	G1
11	48	Unknown	Unknown	G1
12	17	ATGCGAAR	Unknown	S

Table 1.1 Promoter elements (Yeast cell-cycle data)

The results are summarized in Table 1. They found that the majority of the genes in these clusters share common promoter elements, and furthermore, they correspond to known cell-cycle transcription factor binding sites relevant to the time of peak expression.

1.11 CONCLUSIONS

As an increasing number and variety of high-throughput data sets become available, cluster analysis is playing an ever increasing role in the analysis of these biological data. Hierarchical methods have been the primary clustering tool employed to date. The hierarchical algorithms have been mainly applied heuristically to these cluster analysis problems. Also, there is no reason why the clusters of tissues (nor genes)

should belong to a hierarchy such as in the evolution of species. Further, a major limitation of these methods is their inability to determine the number of clusters. Thus there is a need for a model-based approach to this clustering problem. Concerning the clustering of tissue samples, a clustering of, say, some tumors, for example, will reveal whether tumors that have traditionally been lumped together as one type should be divided into a number of distinct subtypes, and whether these subtypes have different prognoses and respond differently to specific therapies. For this clustering problem, we have described the EMMIX-GENE procedure, which is a model-based approach to the clustering of high-dimensional independent observations.

The EMMIX-GENE procedure fits a mixture of multivariate normals without regression structure on the component means and without constraints on the covariance matrices that arise in experimental designs with structure, including replications taken over time. Thus it is not directly applicable to the other clustering problem of grouping the gene-profile vectors as in longitudinal or cross-sectional studies. This problem arises where, say, the interest is to study the changes in gene expression of entire groups of (correlated) genes as a means to finding possible functional relationships among them, the identification of transcription factor binding sites, and the elucidation of biological pathways. The biological rationale underlying the clustering of the gene profiles is the fact that often many coexpressed genes are also coregulated, which is supported both by an immense body of empirical observations and by detailed mechanistic explanation (Boutros and Okey, 2005). However, it has been observed that genes with similar profiles sometimes do not share biological similarity (Clare and King, 2002; Gibbons and Roth, 2002; DeRisi et al., 1997).

Thus clustering does not provide proof of relationships between the genes, but it does provide suggestions that help to direct further research. The idea is we can establish a guilt by association - that is, genes with similar expression patterns are more likely to have similar biological function. For this clustering problem, we have described the EMMIX-WIRE procedure, which provides a unified approach to the clustering of genes with correlated expression levels measured in a wide variety of experimental situations. This procedure is applicable to longitudinal studies with or without replication, for example, time-course experiments by using time as a covariate, and to cross-sectional experiments by using categorical covariates to represent the different experimental classes.

Most clustering algorithms require that one gene be assigned to one cluster, adding an arbitrary element to the analysis. Mixture modeling provides one way to reduce this arbitrariness and to handle the clustering of the borderline cases. It gives a probabilistic or "soft" clustering through the the posterior probabilities of component membership of each gene. An overlapping clustering can be obtained by making a hard assignment of each gene to one or more of the components (clusters) using a threshold on the posterior probabilities of component membership; for example, the jth gene with profile vector $\boldsymbol{y}_j$ belongs to the hth component if its posterior probability of membership of the hth component is greater than some specified threshold c.

References

Banerjee, A., Dhillon, I.S., Ghosh, J., and Sra S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research* **6**, 1345–1382.

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Belitskaya-Levy, I. (2006). A generalized clustering problem, with application to DNA microarrays. *Statistical Applications in Genetics and Molecular Biology* **5**, Article 2.

Booth, J.G., Casella, G., Cooke, J.E.K., and Davis, J.M. (2004). Statistical approaches to analysing microarray data representing periodic biological processes: a case study using the yeast cell cycle, Technical report, Department of Biological Statistics and Computational Biology, Cornell University, 2004.

Bryan, J. (2004). Problems in gene clustering based on gene expression data. *Journal of Multivariate Analysis* **90**, 44–66.

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

Clare, A. and King, R.D. (2002). How well do we understand the clusters in microarray data? *In Silico Biology* **2**, 511–522.

Coleman, D., Dong, X., Hardin, J., Rocke, D.M., and Woodruff, D.L. (1999). Some computational issues in cluster analysis with no a priori metric. *Computational Statistics and Data Analysis* **31**, 1–11.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B* **39**, 1-38.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3**, research0036.1–0036.21

Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* **95**, 14863–14868.

Eisen, M.B. and Brown, P.O. (1999). DNA Arrays for Analysis of Gene Expression. *Methods in Enzymology* **303**, 179-205.

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal* **41**, 578–588.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Friedman, J.H. and Meulman, J.J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society B* **66**, 815–849.

Ganesalingam, S. and McLachlan, G.J. (1978) The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika* **65**, 658-662.

Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Cell Biology* **97**, 12079–12084.

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

Gibbons, F.D. and Roth, F.P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research* **12**, 1574–1581.

Hand, D.J. and Heard, N.A. (2005) Finding groups in gene expression data. *Journal of Biomedicine and Biotechnology* **2005**, 215–225.

Goldstein, H. (1995). *Multilevel Statistical Models (second edition)*. Arnold, London.

Hartigan, J.A. (1975). Statistical theory in clustering, *Journal of Classification* **2**, 63–76.

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

Huang, E., Cheng, S.H., Dressman, H., Pittman, J., Tsou, M.-H., Horng, Ch.-F., Bild, A., Iversen, E.S., Liao, M., Chen, C.-M., West, M., Nevins, J.R., and Huang, A.T. (2003). Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1576–1577.

Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A., and Vingron, M. (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology* **2(1)**, Article 3.

Ideker, T., Thorsson, V., Siegel, A.F., and Hood, L.E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology* **7**, 805–817.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Kettenring, J.R. (2006) The practice of cluster analysis. *Journal of Classification* **23**, 3–30.

Lee, M.L.T., Kuo, F.C., Whitmore, G.A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence

from repetitive cDNA hybridizations, *Proceedings of the National Academy of Sciences USA* **97**, 9834–9838.

Liu, J.S., Zhang, J.L., Palumbo, M.J., and Lawrence, C.E. (2003). Bayesian clustering with variable and transformation selections. In *Bayesian Statistics*, Vol. 7, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West (Eds.). Oxford: Oxford University Press, pp. 249–275.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* **19**, 474–482.

Marriott, F.H.C. (1974) *The Interpretation of Multiple Observations.* Academic Press, London.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

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. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. Dekker, New York.

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.

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

McLachlan, G.J. and Khan, N. (2004). On a resampling approach for tests on the number of clusters with mixture model-based clustering of tissue samples. *Journal of Multivariate Analysis* **90**, 90–105.

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

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

Muro, S., Takemasa, I., Oba, S., Matoba, R., Ueno, N., Maruyama, C., Yamashita, R., Sekimoto, M., Yamamoto, H., Nakamori, S., Monden, M., Ishii, S., and Kato, K. (2003). Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome Biology* **4(5)**, Article R21.

Ng, S.K., Krishnan, T., and McLachlan, G.J. (2004). The EM algorithm. In Gentle, J., Hardle, W., and Mori,Y. (eds), *Handbook of Computational Statistics Vol. 1*. Springer-Verlag, New York, pp. 137–168.

Ng, S.K., McLachlan, G.J., Wang, K., Ben-Tovim, L., and Ng, S.W. (2006a). A mixture model with random-effects components for clustering correlated gene-expression profiles. Submitted.

Ng, S.K., Wang, K., and McLachlan, G.J. (2006b). Multilevel modelling for inference of genetic regulatory networks. In Proceedings of SPIE 2005, *Complex Systems in the International Symposium on Microelectronics, MEMS, and Nanotechnology*, Vol. 6039, A. Bender (Ed.). Bellingham, Washington: International Society for Optical Engineering, pp. 60390S-1–60390S-12.

Pan, W. (2006). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*. To appear.

Pan, W., Lin, J., and Le, C.T. (2002). Model-based cluster analysis of microarray gene-expression data. *Genome Biology* **3**, research0009.1-0009.8.

Parmigiani, G., Garrett, E.S., Irizarry, R.A., and Zeger, S.L. (Eds.) (2003). *The Analysis of Gene Expression Data*. New York: Springer-Verlag.

Pavlidis, P., Li, Q., and Noble, W.S. (2003). The effect of replication on gene expression microarray experiments. *Bioinformatics* **19**, 1620–1627.

Pollard, K. S. and van der Laan, M. J. (2002). Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences* **176**, 99–121.

Reilly, C., Wang, C., and Rutherford, R. (2005). A rapid method for the comparison of cluster analyses. *Statistica Sinica* **15**, 19–33.

Rocke, D.M. and Durbin, B. (2003). Approximate variance-stabilizing transformations for a gene-expression microarray data. *Bioinformatics* **19**, 966–972.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Spang, R. (2003). Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine. *Biosilico* **1**, 64–68.

Speed, T. (Ed.) (2003). *Statistical Analysis of Gene Expression Microarray Data*. Boca Raton, FL: Chapman & Hall/CRC.

Spellman, P., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification

of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

Storey, J.D., Xiao, W. Leek, J.T., Tompkins, R.G., and Davis, R.W. (2005). Significance analysis of time course microarray experiments, *Proceedings of the National Academy of Sciences USA* **102**, 12837–12842.

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