Clustering

G.J. McLachlan, R.W. Bean, and S.K. Ng

Abstract

Clustering techniques are used to arrange genes in some natural order, that is, to organize genes into groups or clusters with similar behaviour across relevant tissue samples (or cell lines). These techniques can also be applied to tissues rather than genes. Methods such as hierarchical agglomerative clustering, k-means clustering, the self-organizing map, and model-based methods have been used. Here we focus on mixtures of normals to provide a model-based clustering of tissue samples and of gene profiles.

Keywords: clustering, hierarchical, dendrogram, tree, k-means clustering, agglomerative, self-organizing map, model-based clustering

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 has 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 has demonstrated its utility in the elucidation of unknown gene function, the validation of gene discoveries, and the interpretation of biological processes; see Alizadeh et al. (2000) and Eisen et al. (1998) for examples.

The main goal of microarray analysis of many diseases, in particular of unclassified cancer, is to identify as yet unclassified cancer subtypes for subsequent validation and prediction, and ultimately to develop individualized prognosis and therapy. Limiting factors include the difficulties of tissue acquisition and the expense of microarray experiments. Thus, often microarray studies attempt to perform a cluster analysis of a small number of tumor samples on the basis of a large number of genes, often resulting in gene-to-sample ratios of approximately 100-fold.

Many researchers have explored the use of clustering techniques to arrange genes in some natural order, that is, to organize genes into groups or clusters with similar behavior across relevant tissue samples (or cell lines). Although a cluster does not automatically correspond to a pathway, it is a reasonable approximation that genes in the same cluster have something to do with each other or are directly involved in the same pathway.

It can be seen 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 in cluster analysis is quite common. In the present context of microarray data, one may be interested in grouping tissues (patients) with similar expression values or in grouping genes on patients with similar types of tumors or similar survival rates.

One of the difficulties of clustering is that the notion of clustering is vague. A useful way to think about the different clustering procedures is in terms of the shape of the clusters produced (Reilly et al., 2005). The majority of the existing clustering methods assume that a similarity measure or metric is known *a priori*; often the Euclidean metric is used. But clearly, it would be more appropriate to use a metric that depends on the shape of the clusters. 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.

Before we proceed to consider the clustering of microarray data, we give a brief account of clustering in a general context. For a more detailed account of cluster analysis, the reader is referred to the many books that either consider or are devoted exclusively to this topic; for example, Everitt (1993), Hartigan (1975a), Hastie et al. (2001, Chapter 14), Kaufman and Rousseeuw (1990), Ripley (1996), and Seber (1984, Chapter 7). A recent review article on clustering is Kettenring (2006).

1.1 Brief Review of Some Clustering Methods

Cluster analysis is concerned with grouping a number (n) of entities into a smaller number (g) of groups on the basis of observations measured on some variables associated with each entity. We let $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{pj})^T$ be the observation or feature vector containing the values of p measurements $y_{1j}, \ldots, y_{pj}$ made on the jth entity ($j = 1, \ldots, n$) to be clustered. These data can be organized as a matrix,

$$\boldsymbol{Y}_{p\times n} = ((y_{vj})); \tag{1}$$

that is, the jth column of $\boldsymbol{Y}_{p\times n}$ is the obervation vector $\boldsymbol{y}_j$.

In discriminant analysis (supervised learning), the data are classified with respect to g known classes and the intent is to form a classifier or prediction rule on the basis of these classified data for assigning an unclassied entity to one of the g classes on the basis of its feature vector. In contrast to discriminant analysis, in cluster analysis (unsupervised learning) there is no prior information on the group structure of the data or, in the case where it is known that the population consists of a number of classes, there are no data of known origin with respect to the classes. The clustering problem falls into two main categories which overlap to some extent (Marriott, 1974):

(i) What is the best way of dividing 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. This is sometimes called dissection or segmentation.

(ii) What is the best way to find a natural subdivision of the entities into groups. Here by natural clusters, it is meant that the clusters can be described as continuous regions of the feature space containing a relatively high density of points, separated from other such regions by regions containing a relatively low density of points (Everitt, 1993). It is therefore intended that natural clusters possess the two intuitive qualities of internal cohesion and external isolation (Cormack, 1971).

Sometimes the distinction between the search for naturally occurring clusters

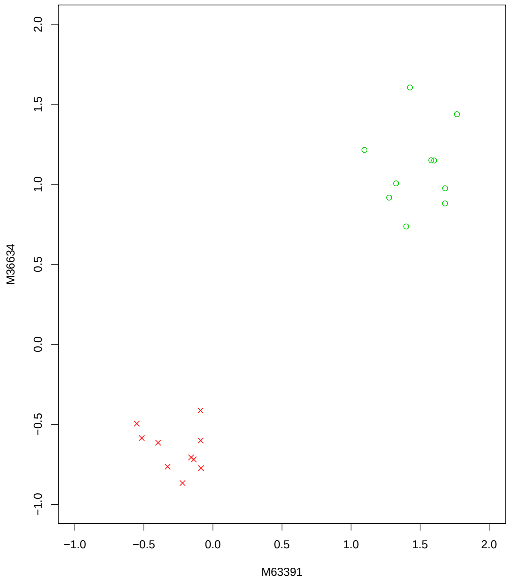


Figure 1: Scatter Plot of the Expression Values of the Two Genes on 10 colon cancer tumours (x) and 10 normal tissues (o).

as in (ii) and other groupings as in (i) is stressed; see, for example, Hand and Heard (2005). But often it is not made, particularly as most methods for finding natural clusters are also useful for segmenting the data. Essentially, all methods of cluster analysis attempt to imitate what the eye-brain does so well in $p = 2$ dimensions. For example, in the scatter plot in Figure **??** of the expression values of two smooth muscle related genes on 10 tumours and ten normal tissues from the colon cancer data of Alon et al. (1999), it is very easy to detect the presence of two clusters of equal size without making the meaning of the term 'cluster' explicit.

Clustering methods can be categorized broadly as being hierarchical or nonhierarchical. With a method in the former category, every cluster obtained at any stage is a merger or split of clusters obtained at the previous stage. Hierarchical methods can be implemented in a so-called agglomerative manner (bottom-up), starting with $g = n$ clusters or in a divisive manner (top-down), starting with the n entities to be clustered as a single cluster. In practice, divisive methods can be computationally prohibitive unless the sample size n is very small. For instance, there are $2^{(n-1)} - 1$ ways of making the first subdivision. Hence hierarchical methods are usually implemented in an agglomerative manner, as to be discussed further in the next section. Chipman and Tibshirani (2006) have proposed a hybrid clustering method that combines the strengths of bottom-up hierarchical clustering with that of top-down clustering. The first method is good at identifying small clusters, but not large ones; the strengths are reversed for top-down clustering.

One of the most popular nonhierarchical methods of clustering is k-means, where "k" refers to the number of clusters to be imposed on the data. It seeks to find $k = g$ clusters that minimize the sum of the squared Euclidean distances between each observation $\boldsymbol{y}_j$ and its respective cluster mean; that is, it seeks to minimize the trace of $\boldsymbol{W}$, tr$\boldsymbol{W}$, where

$$\boldsymbol{W} = \sum_{i=1}^{g} \sum_{j=1}^{n} (\boldsymbol{y}_j - \overline{\boldsymbol{y}}_i)(\boldsymbol{y}_j - \overline{\boldsymbol{y}}_i)^T \tag{2}$$

is the pooled within-cluster sums of squares and products matrix, and

$$\overline{\boldsymbol{y}}_i = \sum_{j=1}^{n} \boldsymbol{y}_j / \sum_{j=1}^{n} z_{ij} \tag{3}$$

is the sample mean of the ith cluster. Here z_{ij} is a zero-one indicator variable that is one or zero, according as $\boldsymbol{y}_j$ belongs or does not belong to the ith cluster $(i = 1, \ldots, g;\ j = 1, \ldots, n)$. It is impossible to consider all partitions of the n

observations into g clusters unless n were very small, since the number of such partitions with nonempty clusters is the Stirling number of the second kind,

$$\frac{1}{n!}\sum_{j=1}^{n}(-1)^{n-1)}\binom{n}{i}n^{g}, \tag{4}$$

which can be approximated by $g^n/g!$; see Kaufman and Rousseeuw (1990). In practice, k-means is therefore implemented by iteratively by moving points between clusters so as to minimize tr$\boldsymbol{W}$. In its simplest form, each observation $\boldsymbol{y}_j$ is assigned to the cluster with the nearest centre (sample mean) and then the centre of the cluster is updated before moving on to the next observation. Often the centres are estimated initially by selecting k points at random from the sample to be clustered.

Other partitioning methods have been developed, including k-medoids (Kaufman and Rousseeuw, 1990), which is similar to k-means, but constrains the each cluster centre to be one of the observations $\boldsymbol{y}_j$. The self-organizing map (Kohonen, 1989) is similar to k-means, but the cluster centres are constrained to lie on a (two-dimensional) lattice. It is well known that k-means tends to lead to spherical clusters since it is predicated on normal clusters with (equal) spherical covariance matrices. One way to achieve elliptical clusters is to seek clusters that minimize the determinant of $\boldsymbol{W}$, $\mid \boldsymbol{W} \mid$, rather than its trace, as in Friedman and Rubin (1967); see also Scott and Symons (1971) who derived this criterion under certain assumptions of normality for the clusters.

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,

$$\boldsymbol{y} \rightarrow \boldsymbol{C}\boldsymbol{y} + \boldsymbol{a}, \tag{5}$$

where $\boldsymbol{C}$ is a nonsingular matrix. If the clustering of a procedure is invariant under (**??**) for only diagonal $\boldsymbol{C}$, then it is invariant under change of measuring units but not rotations. But as commented upon by Hartigan (1975b), this form of invariance is more compelling than affine invariance. The clustering produced by minimization of $| \boldsymbol{W} |$ is affine invariant.

In the statistical and pattern recognition literature in recent times, attention has been focussed on model-based clustering via mixtures of normal densities. With this approach, each observation vector $\boldsymbol{y}_j$ is assumed to have a g-component normal mixture density,

$$f(\boldsymbol{y}_j;\, \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j;\, \boldsymbol{u}_i,\, \boldsymbol{\Sigma}_i), \tag{6}$$

where $\phi(\boldsymbol{y};\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p-variate normal density function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, and the π_i denote the mixing proportions, which are nonnegative and sum to one. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions π_i, the elements of the component means $\boldsymbol{\mu}_i$, and the distinct elements of the component-covariance matrix $\boldsymbol{\Sigma}_i$, and it can be estimated by its maximum likelihood estimate $\hat{\boldsymbol{\Psi}}$ calculated via the EM algorithm; see McLachlan and Basford (1988) and McLachlan and Peel (2000). This approach gives a probabilistic clustering defined in terms of the estimated posterior probabilities of component membership $\tau_i(\boldsymbol{y}_j;\, \hat{\boldsymbol{\Psi}})$, where $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi})$ denotes the posterior probability that the jth feature vector with observed value

$\boldsymbol{y}_j$ belongs to the ith component of the mixture (i=1, ... ,g; j=1, ... ,n). Using Bayes' theorem, it can be expressed as

$$\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}) = \frac{\pi_i \phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \tag{7}$$

It can be seen that with this approach, we can have a "soft" clustering, whereby each observation may partly belong to more than one cluster. An outright clustering can be obtained by assigning y_j to the component to which it has the greatest estimated posterior probability of belonging.

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." One potential drawback with this approach is that normality is assumed for the cluster distributions. However, this assumption would appear to be reasonable for the clustering of microarray data after appropriate normalization.

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or its more robust version in the form of the t density (McLachlan and Peel, 2000) is that the implied clustering is invariant under affine transformations (**??**). Also, in the case where the components of the mixture correspond to externally subpopulations, the unknown parameter vector $\boldsymbol{\Psi}$ can be estimated consistently by a sequence of roots of the likelihood equation. Note that this is not the case if a criterion such as $| \boldsymbol{W} |$ were used.

In the above, we have focussed exclusively on methods that are applicable for the clustering of the observations and the variables considered separately; that is, in the context of clustering microarray data, methods that would be suitable for clustering the tissue samples and the genes considered separately rather than simultaneously. Pollard and van der Laan (2002) have proposed a statistical framework for two-way clustering; see also Getz et al. (2000) and the references therein for earlier approaches on this problem. More recently, Ambroise and Govaert (2006) have reported some results on two-way clustering (biclustering) of tissues and genes. In their work, they obtained similar results to that obtained when the tissues and the genes were clustered separately.

2 Methods

Although biological experiments vary considerably in their design, the data generated by microarray experiments can be viewed as a matrix of expression levels. For M microarray experiments (corresponding to M tissue samples), where we measure the expression levels of N genes in each experiment, the results can be represented by a $N \times M$ matrix. For each tissue, we can consider the expression levels of the N genes, called its *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 $N \times M$ matrix is portrayed in Figure **??**, where each sample represents a separate microarray experiment and generates a set of N expression levels, one for each gene.

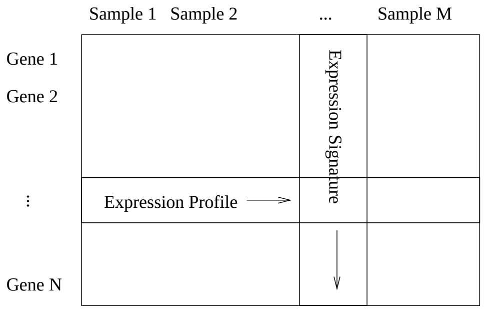


Figure 2: Gene expression data from M microarray experiments represented as a matrix of expression levels with the N rows corresponding to the N genes and the M columns to the M tissue samples.

Against the above background of clustering methods in a general context as given in the previous section, we now consider their application to microarray data, focussing on a model-based approach using normal mixtures. But firstly, we consider the application of hierarchical agglomerative methods given their extensive use for this purpose in bioinformatics.

2.1 Clustering of Tissues: Hierarchical Methods

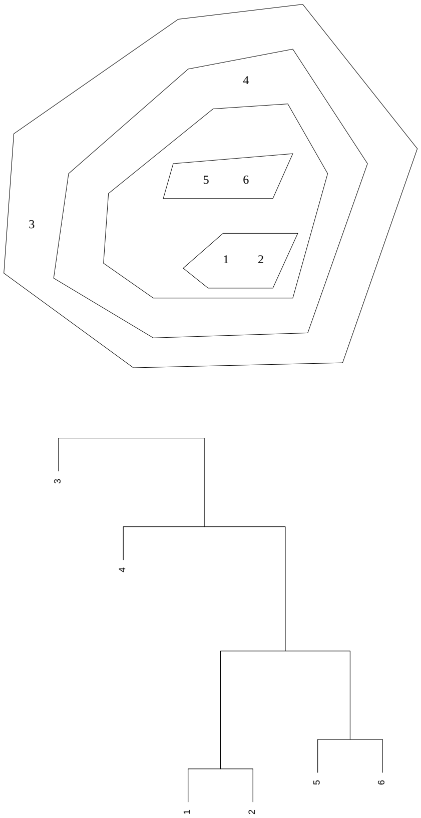
For the clustering of the tissue samples, the microarray data portrayed in Figure **??** are in the form of the matrix (**??**) with $n = M$ and $p = N$, and the observation vector $\boldsymbol{y}_j$ corresponds to the expression signature for the jth tissue sample. In statistics, it is is usual to refer to the entirety of the tissue samples as the sample, whereas the biologists tend to refer to each individual expression signature as a sample, and we follow this practice here.

The commonly used hierarchical agglomerative methods can be applied directly to this matrix to cluster the tissue samples, since they can be implemented by consideration of the matrix of proximities, or equivalently, the distances, between each pair of observations. Thus they require only $O(n^2)$ or at worst $O(n^3)$ calculations, where $n = M$ and the number M of tissue samples is limited usually to being less than 100. The situation would be different with the clustering of the genes as then $n = N$ and the number N of genes could be in the tens of thousands.

In order to compute the pairwise distances between the observations, one needs to select an appropriate distance metric. Metrics that are used include Euclidean distance and the Pearson correlation coefficient, although the latter is equivalent to the former if the observations have been normalized beforehand to have zero means and unit variances. Having selected a distance measure the observations, there is a need to specify a linkage metric between clusters. Some commonly used metrics include single linkage, complete linkage, average linkage, and centroid linkage. With single linkage, the distance between two clusters is defined by the distance between the two nearest observations (one from each cluster), while with complete linkage, the cluster distance is defined in terms of the distance between the two fartherest observations (one from each cluster). Average linkage is defined in terms of the average of the $n_1 n_2$ distances between all possible pairs of observations (one from each cluster), where n_1 and n_2 denote the number of observations in the two clusters in question. For centroid linkage, the distance between two clusters is the distance between the cluster centroids (sample means). Another commonly used method is Ward's (1963) procedure, which joins clusters so as to minimize the within-cluster variance (the trace of $\boldsymbol{W}$). Lance and Williaons (1967) have presented a simple linear system of equations as a unifying framework for these different linkage measures. Eisen et al. (1998) were the first to apply cluster analysis to microarray data, using average linkage with a correlation-based metric. The nested clusters produced by an hierarchical method of clustering can be portrayed in a tree diagram, in which the extremities (usually shown at the bottom) represent the indiviudal observations, and the branching of the tree gives the order of joining together. The height at which clusters of points are joined corresponds to the distance between the clusters. However, it is not clear in general how to choose the number of clusters.

To illustrate hierarchical agglomerative clustering, we use nested polygons in Figure **??** to show the clusters obtained by applying it to six bivariate points, using single-linkage with Euclidean distance as the distance measure. It can be seen that the cluster of observations 3 and and 4 is considerably closer to the cluster of 1 and 6 than what observation 5 is.

There is no reason why the clusters should be hierarchical for microarray data. 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



[Nested polygons]

[Dendrogram]

Figure 3: Hierarchical agglomerative clustering with different representations

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. In the sequel, we therefore focus on nonhierarchical methods of clustering. As advocated by Marriott (1974, Page 67), "it is better to consider the clustering problem *ab initio*, without imposing any conditions."

2.2 Clustering of Tissues: Normal Mixtures

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

Banfield and Raftery (1993) introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. But if p is large relative to the sample size n, it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it were possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when p is large relative to n.

Hence, in fitting normal mixture models with unrestricted component-covariance matrices to high-dimensional data, we need to consider first some form of dimension reduction and/or some form of regularization. A common approach to reducing the the number of dimensions is to perform a principal component analysis (PCA). However, the latter provides only a global linear model for the representation of the data in a lower-dimensional subspace. Thus it has limited scope in revealing group structure in a data set. A global nonlinear approach can be obtained by postulating a finite mixture of linear (factor) submodels for the distribution of the full observation vector $\boldsymbol{y}_j$ given a relatively small number of (unobservable) factors. That is, we can provide a local dimensionality reduction method by a mixture of factor analyzers model, which is given by (**??**) by imposing on the component-covariance matrices $\boldsymbol{\Sigma}_i$, the constraint

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g), \tag{8}$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1, \ldots, g)$. We can think of the use of this mixture of factor analyzers model as being purely a method of regularization, but in the present context, it might be possible to make a case for it being a reasonable model for the correlation structure between the genes.

The EMMIX-GENE program of McLachlan et al. (2002) has been designed for the clustering of tissue samples via mixtures of factor analyzers. 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 (Pollard and van der Laan, 2002; Friedman and Meulman, 2004). 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. The usefulness of a given gene to the clustering process can be assessed formally by a test of the null hypothesis that it has a single component normal distribution over the tissue samples. A faster but *ad hoc* way is to make this decision on the basis of the interquartile range. Even after this step has been completed, there may still remain too many genes. Thus there is a second step in EMMIX-GENE in which the retained gene profiles are clustered (after standardization) into a number of groups on the basis of Euclidean distance so that genes with similar profiles are put into the same group. In general, care has to be taken with the scaling of variables before clustering of the observations, as the nature of the variables can be intrinsically different. In the present context there is not this problem, as the variables (gene expressions) are measured on the same scale. Also, as noted above, the clustering of the observations (tissues) via normal mixture models is invariant under changes in scale and location. The clustering of the tissue samples can be carried out on the basis of the groups considered individually using some or all of the genes within a group or collectively. For the latter, we can replace each group by a representative (a metagene) such as the sample mean as in the EMMIX-GENE procedure.

To illustrate this approach, we applied the EMMIX-GENE procedure to the colon cancer data of Alon et al. (1999). It consists of $n = 2000$ genes and $p = 62$ columns denoting 40 tumours and 22 normal tissues. After applying select-genes to this set, there were 446 genes remaining in the set. The remaining genes were then clustered into 20 groups, which were ranked on the basis of

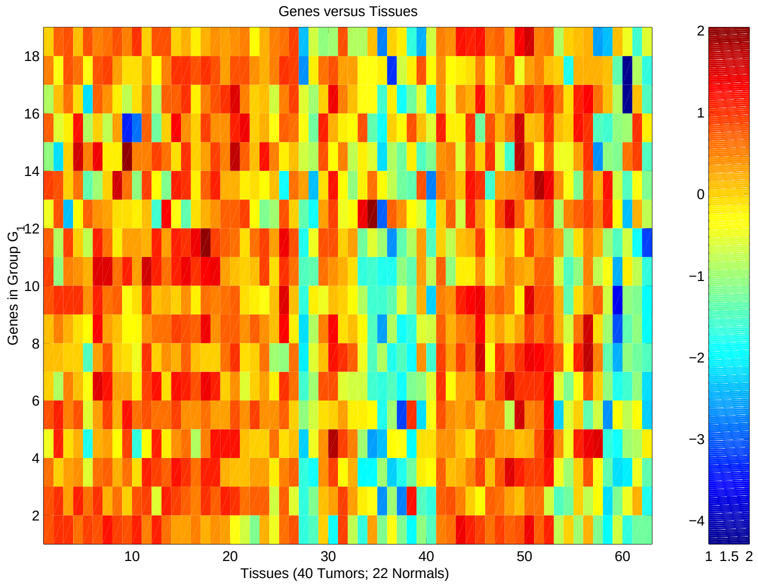


Figure 4: Heat map of 18 genes in group G_1 on 40 tumor and 22 normal tissues in Alon data.

$-2 \log \lambda$, where λ is the likelihood ratio statistics for testing $g = 1$ versus $g = 2$ components in the mixture model. The heat map of the second ranked group G_2 is shown in Figure **??**. The clustering of the tissues on the basis of the 24 genes in G_2 resulted in a partition of the tissues in which one cluster contains 37 tumors (1–29, 31–32, 34–35, 37–40) and 3 normals (48, 58, 60), and the other cluster contains 3 tumors (30, 33, 36) and 19 normals (41–47, 49–57, 59, 61–62). This corresponds to an error rate of 6 out of 62 tissues compared to the "true" classification given by Alon et al. (2002). For further details about the results of the cluster-tissues procedure on this data set, see McLachlan et al. (2002).

2.3 Clustering of Gene Profiles

In order to cluster gene profiles, it might seem possible just to interchange rows and columns in the data matrix (**??**). But with most applications of cluster analysis in practice it is assumed that

(a) there are no replications on any particular entity specifically identified as such;

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

These assumptions should hold for the clustering of the tissue samples, although the tissue samples have been known to be correlated for different tissues due to flawed experiemental conditions. However, condition (b) will not hold for the clustering of gene profiles, since not all the genes are independently distributed, and condition (a) will generally not hold either as the gene profiles may be measured over time or on technical replicates. While this correlated structure can be incorporated into the normal mixture model (**??**) by appropriate specification of the component-covariance matrices Σ_i, it is difficult to fit the model under such specifications. For example, the M-step may not exist in closed form. Accordingly, we now consider the EMMIX-WIRE model of Ng et al. (2006), who adopt conditionally a mixture of linear mixed models to specify this correlation structure among the tissues samples and to allow for correlations among the genes. It also enables covariate information to be incorporated into the clustering process.

For a gene microarray experiment with repeated measurements, we have for the jth gene ($j = 1, \dots, n$), a feature vector (profile vector) $\boldsymbol{y}_j = (\boldsymbol{y}_{1j}^T, \dots, \boldsymbol{y}_{tj}^T)^T$, where t is the number of distinct tissues in the experiment and

$$\boldsymbol{y}_{lj} = (y_{l1j}, \dots, y_{lrj})^T \qquad (l = 1, \dots, t)$$

contains the r replications on the jth gene from the lth tissue. Conditional on its membership of the ith component of the mixture, the EMMIX-WIRE procedure assumes that $\boldsymbol{y}_j$ follows a linear mixed-effects model (LMM),

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{U}\boldsymbol{b}_{ij} + \boldsymbol{V}\boldsymbol{c}_i + \boldsymbol{\epsilon}_{ij}, \tag{9}$$

where the elements of $\boldsymbol{\beta}_i$ (a t-dimensional vector) are fixed effects (unknown constants ($i = 1, \dots, g$). In (**??**), $\boldsymbol{b}_{ij}$ (a q_b-dimensional vector) and $\boldsymbol{c}_i$ (a q_c-dimensional vector) represent the unobservable gene- and tissue-specific random effects, respectively, conditional on membership of the ith cluster. These random effects represent the variation due to the heterogeneity of genes and tissues (corresponding to $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}^T, \dots, \boldsymbol{b}_{in}^T)^T$ and $\boldsymbol{c}_i$, respectively). The random effects $\boldsymbol{b}_i$ and $\boldsymbol{c}_i$, and the measurement error vector $(\boldsymbol{\epsilon}_{i1}^T, \dots, \boldsymbol{\epsilon}_{in}^T)^T$ are assumed to be mutually independent, where $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are known design matrices of the corresponding fixed or random effects.

With the LMM, the distributions of $\boldsymbol{b}_{ij}$ and $\boldsymbol{c}_i$ are taken, respectively, to be multivariate normal $N_{q_b}(\boldsymbol{0}, \theta_{bi}\boldsymbol{I}_{q_b})$ and $N_{q_c}(\boldsymbol{0}, \theta_{ci}\boldsymbol{I}_{q_c})$, where $\boldsymbol{I}_{q_b}$ and $\boldsymbol{I}_{q_c}$ are identity matrices with dimensions being specified by the subscripts. The measurement error vector $\boldsymbol{\epsilon}_{ij}$ is also taken to be multivariate normal $N_m(\boldsymbol{0}, \boldsymbol{A}_i)$, where $\boldsymbol{A}_i = \text{diag}(\boldsymbol{H}\boldsymbol{\phi}_i)$ is a diagonal matrix constructed from the vector $(\boldsymbol{H}\boldsymbol{\phi}_i)$

with $\boldsymbol{\phi}_i = (\sigma_{i1}^2, \dots, \sigma_{iq_e}^2)^T$ and $\boldsymbol{H}$ is a known $m \times q_e$ zero-one design matrix. That is, we allow the ith component-variance to be different among the m microarray experiments.

The vector $\boldsymbol{\Psi}$ of unknown parameters can be obtained by maximum likelihood via the EM algorithm, proceeding conditionally on the tissue-specific random effects $\boldsymbol{c}_i$. The E- and M-steps can be implemented in closed form. In particular, an approximation to the E-step by carrying out time-consuming Monte Carlo methods is not required. A probabilistic or an outright clustering of the genes into g components can be obtained, based on the estimated posterior probabilities of component membership given the profile vectors and the estimated tissue-specific random effects $\hat{\boldsymbol{c}}_i \, (i = 1, \dots, g)$.

To illustrate this method, we report here an example from Ng et al. (2006) who used it to cluster some time course data from the yeast cell-cycle study of Spellman et al. (1998). The data consist of the expression levels of 612 genes for the yeast cells at $M = 18$ time points. With reference to (**??**), the design matrix $\boldsymbol{X}$ was taken be an 18×2 matrix with the $(l+1)$th row $(l = 0, \dots, 17)$

$$(\cos(2\pi(7l)/\omega + \Phi) \quad \sin(2\pi(7l)/\omega + \Phi)), \tag{10}$$

where the period of the cell cycle ω was taken to be 53 and the phase offset Φ was set to zero. The design matrices for the random effects parts were specified as $\boldsymbol{U} = \mathbf{1_{18}}$ and $\boldsymbol{V} = \boldsymbol{I}_{18}$. That is, it is assumed that there exists random gene effects b_{ij} with $q_b = 1$ and random temporal effects $(c_{i1}, \dots, c_{iq_c})$ with $q_c = m = 18$. The latter introduce dependence among expression levels within the same cluster obtained at the same time point. Also, $\boldsymbol{H} = \mathbf{1_{18}}$ and $\boldsymbol{\phi}_i = \sigma_i^2$

($q_e = 1$) so that the component variances are common among the $m = 18$ experiments.

The number of components g was determined using BIC for model selection. It indicated here that there are twelve clusters. The clustering results for $g = 12$ are given in Figure **??**, where the expression profiles for genes in each cluster are presented.

3 Notes

We have described the EMMIX-GENE and EMMIX-WIRE procedures for the clustering of tissue samples (expression signatures) and of gene profiles, respectively. Both procedures are implemented by maximum likelihood via the EM algorithm. The EMMIX-GENE procedure clusters the data by fitting a normal mixture model to the signatures with a factor analytic constraint on each component-covariance matrix after an initial screening and then clustering of the selected genes into metagenes. Here the latter clustering of the genes is not an end in itself but rather a way of reducing the number of genes (variables) to be used in the clustering of the tissue samples (observations). The EMMIX-WIRE procedure clusters gene-expression profiles by postulating conditionally a mixture of linear mixed-effects models for the expression profiles. The unconditional density of the expression profile is given by averaging this normal mixture model over its gene- and tissue-specific random effects. This density does not exist in closed form, but the M-step of the EM algorithm is able to be

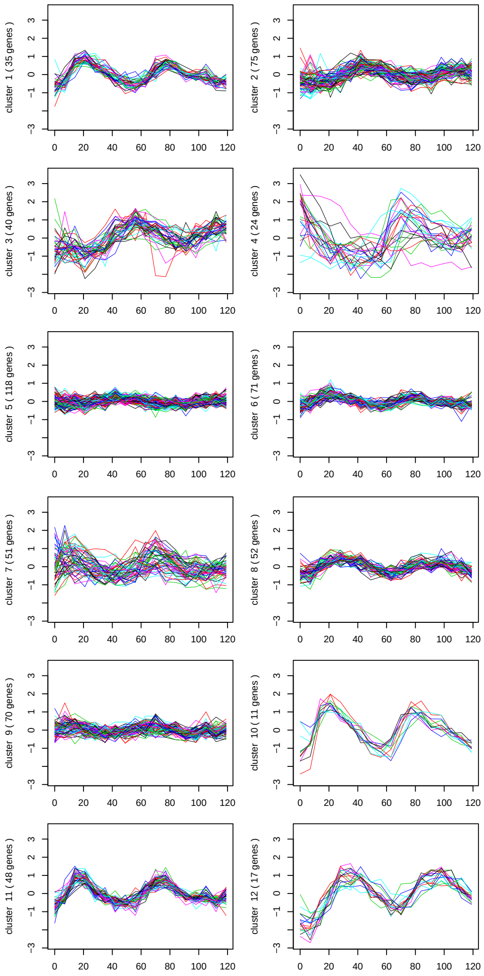


Figure 5: Clustering Results for Spellman Yeast Cell Cycle Data. For all the plots, the x-axis is the time point and the y-axis is the gene-expression level.

carried exactly since the complete-data log likelihood can be written down in closed form.

For both procedures, as with other partitional clustering methods, the number of clusters g needs to be specified at the outset. As both procedures are model-based, we can make a choice as to an appropriate value of g by consideration of the likelihood function. In the absence of any prior information as to the number of clusters present in the data, we monitor the increase in the log likelihood function as the vlaue of g increases. At any stage, the choice of $g = g_o$ versus $g = g_1$, for instance $g_1 = g_o$, can be made by either performing the likelihood ratio test or by using some information-based criterion, such as BIC (Bayesian information criterion). Unfortunately, regularity conditions do not hold for the likelihood ratio test statistic λ to have its usual null distribution of chi-squared with degrees of freedom equal to the difference d in the number of parameters for $g = g_1$ and $g = g_o$ components in the mixture models. One way to proceed is to use a resampling approach as in McLachlan (1987). Alternatively, one can apply BIC, which leads to the selection of $g = g_1$ over $g = g_o$ if $-2 \log \lambda$ is greater than $d \log(n)$. The value of d is obvious in applications of EMMIX-GENE, but is not so clear with applications of EMMIX-WIRE, due to the presence of random effects terms (Ng et al., 2006).

4 References

Aitkin, M., Anderson, D., and Hinde, J. (1981) Statistical modelling of data on teaching styles (with discussion). J. R. Statist. Soc. A 144, 419–461.

Alizadeh, A., Eisen, M.B., Davis, R.E., et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 50 3–511.

Ambroise, C. and Govaert, G. (2006) Model based hierarchical clustering. Unpublished manuscript.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences USA 96, 6745–6750.

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* Statist. Appl. Genetics Mol. Biol. 5, Article 2.

Chipman, H. and Tibshirani, R. (2006) Hybrid hierarchical clustering with applications to microarray data. Biostatistics 7, 286–301.

Coleman, D., Dong, X.P., Hardin, J., Rocke, D.M., Woodruff, D.L. (1999) Some computational issues in cluster analysis with no a priori metric. Comput. Statist. Data Anal. 31, 1-11.

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., 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.

Everitt, B.S. (1993) Cluster Analysis, 3rd edition. Edward Arnold, London.

Friedman, H.P. and Rubin, J. (1967) On some invariant criteria for grouping data. J. Amer. Statist. Assoc. 62, 1159–1178.

Friedman, J.H. and Meulman, J.J. (2004) Clustering objects on subsets of attributes (with discussion). J. R. Statist. Soc. B 66, 815–849.

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

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

Hand, D.J. and Heard, N.A. (2005) Finding groups in gene expression data. J. Biomed. Biotech. 2005, 215–225.

Hartigan, J.A. (1975a) Clustering Algorithms. Wiley, New York.

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

Hastie, T., Tibshirani, R.J., and Friedman, J.H. (2001) The Elements of Statistical Learning. Springer-Verlag, New York.

Kaufman, L. and Rousseeuw, P. (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York.

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

Kohonen, T. (1989) Self-organization and Associative Memory, 3rd edition. Springer-Verlag, Berlin.

Lance, G.N. and Williams, W.T. (1967) A generalized theory of classificatory sorting strategies: I. Hierarchical systems. Comp. J. 9, 373–380.

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.

McLachlan, G.J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl. Statist. 36, 318–324.

McLachlan, G.J. and Basford, K.E. (1988) Mixture Models: Inference and Applications to Clustering. Marcel 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. 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.

Ng, S. K., McLachlan G. J., Wang, K., Ben-Tovim Jones L. and Ng S.-W. (2006) A mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics 22, 1745–1752.

Pollard, K.S. and van der Laan, M.J. (2002) Statistical inference for simultaneous clustering of gene expression data. Math. Biosci. 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.

Ripley, B.D. (1996) Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, UK.

Scott, A.J. and Symons, M.J. (1971) Clustering methods based on likelihood ratio criteria. Biometrics 27, 387–397.

Seber, G.A.F. (1984) Multivariate Observations. Wiley, New York.

Spellman, P., Sherlock, G., Zhang, M.Q., et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Mol. Biol. Cell 9, 3273-3297.

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.