Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Journal of Multivariate Analysis 90 (2004) 90–105

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

On a resampling approach for tests on the number of clusters with mixture model-based clustering of tissue samples

G.J. McLachlan[a,b,*,1] and N. Khan[a]

[a] Department of Mathematics, University of Queensland, St. Lucia, Queensland, Brisbane 4072, Australia
[b] Institute for Molecular BioScience, University of Queensland, St. Lucia, Queensland, Brisbane 4072, Australia

Received 1 April 2003

Abstract

We consider the problem of assessing the number of clusters in a limited number of tissue samples containing gene expressions for possibly several thousands of genes. It is proposed to use a normal mixture model-based approach to the clustering of the tissue samples. One advantage of this approach is that the question on the number of clusters in the data can be formulated in terms of a test on the smallest number of components in the mixture model compatible with the data. This test can be carried out on the basis of the likelihood ratio test statistic, using resampling to assess its null distribution. The effectiveness of this approach is demonstrated on simulated data and on some microarray datasets, as considered previously in the bioinformatics literature.

AMS 2000 subject classifications: 65C60

Keywords: Microarray gene expression data; Mixture models; Clustering of tissue samples; Tests on number of clusters; Likelihood ratio statistic; Resampling approach

[*] Corresponding author. Fax: +61-7-33651477.
E-mail address: gjm@maths.uq.edu.au (G.J. McLachlan).

[1] Supported by a Grant of the Australian Research Council.

doi:10.1016/j.jmva.2004.02.002

1. Introduction

The analysis of gene expression microarray data using clustering techniques has an important role to play in the discovery, validation, and understanding of various classes and subclasses of cancer; see, for example, [2,4,6,7,9,15,17,20,28,30,40,43,46] among several others. This paper considers a mixture model-based approach to the clustering of tissue samples of a very large number of genes from microarray experiments. As commented by Yeung et al. [47], "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". They have advocated a model-based approach to clustering by adopting a finite mixture model for the distribution of each observation. In their study, they were concerned with the clustering of the genes on the basis of the tissue samples. Here we consider the problem of clustering the tissues on the basis of the genes, which is a more challenging problem to consider in a mixture model framework, since the number of observations to be clustered (the tissue samples) is typically small relative to the number of genes in each tissue sample.

More specifically, we consider the cluster analysis of M tissue samples, each containing N genes from a microarray experiment. These microarray data can be represented in the form of a $N \times M$ data matrix $\boldsymbol{A}$ whose ith row contains the expression levels for the ith gene in the M tissue samples. Typically, M is no more than 100, while the number of genes N is of the order of 10^4. The expression levels are taken to be the measured (absolute) intensities for oligonucleotide microarrays and the ratios of the intensities for the Cy5-channel (red) images and Cy3-channel (green) images for cDNA microarrays; see, for example, [13]. 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, [24,25,38,41,44].

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 $\boldsymbol{y}_j$,

$$n \gg p. \tag{1}$$

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 this paper we are to consider the cluster analysis of the M tissue samples on the basis of the N genes. For this problem, we have $n = M$ and $p = N$, and so the sample size n will be typically small relative to the dimension p, thus causing estimation problems under the normal mixture model. This is because the g-component normal mixture model (4) with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. It therefore cannot be fitted directly to the tissues on the basis of all the $p = N$ genes.

One way to handle this dimensionality problem is to ignore the correlations between the genes and to cluster the tissue samples by fitting mixtures of normal component distributions with diagonal covariance matrices. This is essentially equivalent to using the k-means clustering procedure if we take the common diagonal covariance matrix to be a multiple of the $p \times p$ identity matrix $\boldsymbol{I}_p$ and impose the additional restriction of equal mixing proportions. However, this leads to spherical clusters whereas, in practice, the clusters tend to be elliptical and not necessarily parallel to the axes in the feature space.

This led McLachlan et al. [30] to develop the software called EMMIX-GENE, which enables elliptical clusters of arbitrary orientation to be imposed on the tissue samples. The EMMIX-GENE program handles this high-dimensional problem 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 the given component. Before the actual fitting of the mixtures of factor analyzers, it is recommended that the number of genes be reduced to a manageable number by firstly 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. Then, secondly, the retained genes are clustered into groups on the basis of Euclidean distance so that highly correlated genes are clustered into the same group. The mixtures of factor analyzers model can then be applied 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.

2. Mixture model-based clustering

2.1. Normal mixtures

With a mixture model-based approach to clustering, it is assumed that the data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ to be clustered are from a mixture of an initially specified number g of groups in various proportions. That is, each data point $\boldsymbol{y}_j$ is taken to be a realization of a p-dimensional random vector with mixture density

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j), \tag{2}$$

where the g components correspond to the g groups. In (2), the $f_i(\boldsymbol{y}_j)$ are densities and the π_i are nonnegative quantities (the mixing proportions) that sum to one.

On specifying a parametric form $f_i(\boldsymbol{y}_j;\, \boldsymbol{\theta}_i)$ for each component density, we can fit this parametric mixture model

$$f(\boldsymbol{y}_j;\, \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \tag{3}$$

by maximum likelihood via the expectation-maximization (EM) algorithm of Dempster et al. [12]; see also [31]. Here

$$\boldsymbol{\Psi} = (\boldsymbol{\xi}^T, \pi_1, \ldots, \pi_{g-1})^T$$

is the vector of unknown parameters, where $\boldsymbol{\xi}$ consists of the elements of the $\boldsymbol{\theta}_i$ known a priori to be distinct. Once the mixture model has been fitted, a probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data. An outright assignment of the data into g clusters is achieved by assigning each data point to the component to which it has the highest estimated posterior probability of belonging.

We consider here the use of mixture models with normal components. That is, the ith component density for the jth observation $\boldsymbol{y}_j$ is specified as

$$f_i(\boldsymbol{y}_j;\, \boldsymbol{\theta}_i) = \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{4}$$

where $\phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p-variate normal density function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. For unrestricted component-covariance matrices $\boldsymbol{\Sigma}_i$, care has to be taken that the EM algorithm has converged to a local maximizer, since the likelihood is unbounded. Also, 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.

If need be, the normal mixture model can be made less sensitive to outlying observations by using t-component densities, as in [33,34,39]. With this t-mixture model-based approach, the normal distribution for each component in the mixture is embedded in a wider class of elliptically symmetric distributions with an additional parameter called the degrees of freedom. The advantage of the t-mixture model is that, although the number of outliers needed for breakdown is almost the same as with the normal mixture model, the outliers have to be much larger [21].

Usually, there is no a priori metric (or equivalently a user-defined distance matrix) for a cluster analysis. In this case, 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; see [11] on the desirability of this invariance in estimation and clustering.

2.2. Mixtures of factor analyzers

The g-component normal mixture model with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. A common approach to reducing the number of dimensions is to perform a principal component analysis (PCA) and then to perform the cluster analysis on the basis of the first few leading principal components as, for example, in [18]. But as is well known, projections of the feature data $\boldsymbol{y}_j$ onto the first few principal axes are not always useful in portraying the group structure.

One approach for reducing the number of unknown parameters in the forms for the component-covariance matrices $\boldsymbol{\Sigma}_i$ is to adopt the mixtures of factor analyzers model, as considered in [34,35]. With the mixture of factor analyzers model, the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g), \tag{5}$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix. Unlike the PCA model, the factor analysis model (5) enjoys a powerful invariance property: changes in the scales of the feature variables in $\boldsymbol{y}_j$, appear only as scale changes in the appropriate rows of the matrix $\boldsymbol{B}_i$ of factor loadings.

The elements of the diagonal matrix $\boldsymbol{D}_i$ (the uniquenesses) will be close to zero if effectively not more than q observations are unequivocally assigned to the ith component of the mixture in terms of the fitted posterior probabilities of component membership. This will lead to spikes or near singularities in the likelihood [37]. One way to avoid this is to impose the condition of a common value $\boldsymbol{D}$ for the $\boldsymbol{D}_i$,

$$\boldsymbol{D}_i = \boldsymbol{D} \quad (i = 1, \ldots, g). \tag{6}$$

In our experience with microarray data sets, we have found that the choice of the number of factors q is not crucial in the clustering of the tissue samples. A formal test for q can be undertaken using the likelihood ratio λ, as regularity conditions hold for this test conducted at a given value for the number of components g. For the null hypothesis that $H_0 : q = q_0$ versus the alternative $H_1 : q = q_0 + 1$, the statistic $-2 \log \lambda$ is asymptotically chi-squared with $d = g(p - q_0)$ degrees of freedom. However, in situations where n is not large relative to the number of unknown parameters, we prefer the use of the BIC criterion of Schwarz [42]. Applied in this context, it means that twice the increase in the log likelihood $(-2 \log \lambda)$ has to be greater than $d \log n$ for the null hypothesis to be rejected.

3. EMMIX-GENE software

The EMMIX-GENE software is an extension of the EMMIX program, as developed by McLachlan et al. [36] for standard clustering problems. It has the facility to fit mixtures of factor analyzers to the tissue samples. 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. Therefore, the EMMIX-GENE program has two optional stages before the final stage of clustering the tissues. First, the genes are screened on an individual basis to eliminate those which have little variation across all the tissue samples in terms of the likelihood ratio test statistic. Second, the retained genes are clustered into groups on the basis of Euclidean distance so that highly correlated genes are clustered into the same group. The Euclidean distance between any two genes is equal to $2(n - 1)(1 - r)$, where r denotes the sample correlation between them, since it is assumed that the genes (that is, the rows of the data matrix $\boldsymbol{A}$) have been standardized to have mean zero and unit standard deviation. This normalization of the gene-expression profiles does not affect the clustering of the tissues because, as discussed in Section 2, the clustering algorithm used here is invariant under affine transformations. The third and final stage of the EMMIX-GENE approach concerns the clustering of the tissues by fitting mixtures of factor analyzers. It can be undertaken on the basis of (i) all or a selected subset of the available genes, (ii) all or some of the gene-group means, or (iii) all or some of the genes within a specified gene group.

4. Likelihood ratio test for the number of clusters

With a mixture model-based approach to clustering, the question of how many clusters there are can be considered in terms of the number of components of the mixture model being used. It is sensible in practice to approach the latter question of the number of components g in a mixture model in terms of an assessment of the smallest number of components compatible with the data, as discussed in [34, Section 6.1]. In the above formulation of the mixture model for clustering, there is a one-to-one correspondence between the mixture components and the groups. In those cases where the underlying population consists of groups in which the feature vector is unable to be modelled by a single component (normal) distribution but needs a normal mixture, the components in the fitted g-component normal mixture model and the consequent clusters will correspond to g subgroups rather than to the smaller number of actual groups represented in the data.

A guide to the final choice of g can be obtained from monitoring the increase in the log likelihood as g is increased from a single component. Unfortunately, it is difficult to carry out formal tests at any stage of this sequential process for the need of an additional component, since, as is well known, regularity conditions fail to hold for the likelihood ratio test statistic $-2\log\lambda$ to have its usual asymptotic null distribution of chi-squared with degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses. Here λ denotes the likelihood ratio; see [34, Section 6.9]. A formal test of the null hypothesis $H_0 : g = g_0$ versus the alternative $H_1 : g = g_1$ $(g_1 > g_0)$ can be undertaken as in [29]. He proposed a resampling approach to the assessment of the P-value of the likelihood ratio test statistic in testing

$$H_0 : g = g_0 \text{ vs. } \quad H_1 : g = g_1 \tag{7}$$

for a specified value of g_0. Previously, Aitkin et al. [1] had adopted a resampling approach in the context of a latent class analysis; see also [5,23].

Bootstrap samples are generated from the mixture model fitted under the null hypothesis of g_0 components. That is, the bootstrap samples are generated from the mixture model with the vector $\boldsymbol{\Psi}$ of unknown parameters replaced by its maximum likelihood estimate (MLE) $\hat{\boldsymbol{\Psi}}_{g_0}$ computed by consideration of the log likelihood formed from the original data under H_0. The value of $-2\log\lambda$ is computed for each bootstrap sample after fitting mixture models for $g = g_0$ and g_1 in turn to it. The process is repeated independently a number of times B, and the replicated values of $-2\log\lambda$ formed from the successive bootstrap samples provide an assessment of the bootstrap, and hence of the true, null distribution of $-2\log\lambda$. It enables an approximation to be made to the achieved level of significance P corresponding to the value of $-2\log\lambda$ evaluated from the original sample. The rth-order statistic of the B bootstrap replications can be used to estimate the quantile of order $r/(B+1)$. A preferable alternative would be to use the rth-order statistic as an estimate of the quantile of order $(3r-1)/(3B+1)$ [22].

In general, the use of the estimate $\hat{\boldsymbol{\Psi}}_{g_0}$, in place of the unknown value of $\boldsymbol{\Psi}$ under the null hypothesis, will affect the accuracy of the P-values assessed on the basis of the bootstrap replications of $-2\log\lambda$. McLachlan and Peel [32] performed some simulations to demonstrate this effect. They observed that there was a tendency for the resampling approach using bootstrap replications to underestimate the upper percentiles of the null distribution of $-2\log\lambda$, and hence underestimate the P-value of tests based on this statistic.

5. Some other methods for assessing the number of clusters

Before we proceed to report some results on the likelihood ratio test with resampling for assessing the number of clusters, we briefly consider some other methods for this problem.

Dudoit and Fridlyand [13] proposed a prediction-based method for assessing the number of clusters in the data, which they called Clest. It is concerned with the reproducibility or predictability of the clusters. For a fixed number of clusters g, it proceeds by repeatedly dividing the original sample into two sets, a training or learning set $S_{\mathrm{L},b}$ and a test set $S_{\mathrm{T},b}$ on a given replication b. A clustering of $S_{\mathrm{L},b}$ is obtained and a classifier is found on the basis of this clustering as if the cluster labels were the true class labels. This classifier is then applied to the test set $S_{\mathrm{T},b}$ and the predicted group labels are compared using some external index a_b. This procedure is repeated B times to give $a_1,\dots,a_B$ and their median m_g. The null distribution of m_g is approximated by the bootstrap under the uniformity hypothesis whereby the data are sampled from a uniform distribution in p-dimensional space. If $m^*_{g,1},\dots,m^*_{g,B_o}$ denote the B_o bootstrap values corresponding to m_g so obtained, we let $\bar{m}^*_g$ denote their sample mean and ω^*_g is taken to be the proportion of these B_o bootstrap samples that are at least as large as m_g (the assessed P-value). Finally, let $d^*_g = m_g - \bar{m}^*_g$.

To complete the definition of the Clest procedure, we need the set J, which is defined as

$$J = \{2 \leqslant g \leqslant g_{\max};\, \omega^*_g \leqslant \omega_{\max},\, d^*_g \geqslant d_{\min}\},$$

where $g_{\max}$ is the maximum value of g to be considered and $\omega_{\max}$ and $d_{\min}$ are preset thresholds. The ad hoc choice in [13] for $\omega_{\max}$ and $d_{\min}$ was 0.05 each. If this set J is empty, the number of clusters is estimated as one ($\hat{g}=1$). Otherwise, let the number of clusters is estimated by

$$\hat{g} = \arg\max_{g} d_g^*.$$

Dudoit and Fridlyand [13] applied their test procedure using the partitioning around medoids (PAM) method of Kaufman and Rousseeuw [26], a linear normal-based classifier with diagonal group covariance matrices, and the external index of Fowlkes and Mallows [16]. They compared the performance of their Clest procedure with six other methods using simulated data and gene-expression data from four published cancer microarray studies. The six methods were the silhouette criterion of Kaufman and Rousseeuw [26], the gap/gapPC statistics of Tibshirani et al. [45], and the criteria proposed by Caliński and Harabasz [10], Krzanowski and Lai [27], and Hartigan [19].

6. Simulation results

We now report the results of some simulation experiments performed to compare the likelihood ratio test statistic (LRT) under the normal mixture model with the Clest procedure [13] for the choice of number of clusters. We used the same eight population models as adopted by Dudoit and Fridlyand [13] to compare their Clest procedure with six other criteria, using the same number of replications (50) per model. From their simulations, Dudoit and Fridlyand [13] concluded that Clest was the most robust and accurate. Hence we extracted only the performance of the Clest procedure from [13] to compare with the LRT in Table 1 obtained from our simulations.

The eight models can be described briefly as follows, where $\boldsymbol{Y}_{ij}$ $(j = 1, \ldots, n_i)$ denote the n_i observations generated independently in group G_i $(i = 1, \ldots, g)$.

Model 1: $(g = 1, p = 10, n = 200)$, where $\boldsymbol{Y}_{ij}$ is from the uniform distribution over the unit hypercube in 10 dimensions.

Model 2: $(g = 3, p = 2, n_1 = 25, n_2 = 25, n_3 = 50)$, where

$$\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{I}_2)$$

and where

$$\boldsymbol{\mu}_1 = (0,\ 0)^T, \quad \boldsymbol{\mu}_2 = (0, 5)^T \quad \text{and} \quad \boldsymbol{\mu}_3 = (5, -3)^T.$$

Model 3: $(g = 4, p = 10, n_i = 25$ or 50 with probability 0.5 each), where

$$\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{I}_{10})$$

and where

$$\boldsymbol{\mu}_i = (\boldsymbol{w}_i^T, \boldsymbol{0}_7^T)^T$$

and $\boldsymbol{w}_i$ is a realization of the random variable $\boldsymbol{W}_i$ distributed as

$$\boldsymbol{W}_i \sim N(\boldsymbol{0}_3, 25\boldsymbol{I}_3).$$

Table 1
Estimating the number of clusters in simulated data

Model	Method	Number of clusters			
		1*	2	3	4
1	Clest	48	2	0	0
1	LRT	50	0	0	0
		1	2	3*	4
2	Clest	0	1	49	0
2	LRT	0	0	49	0
		1	2	3	4*
3	Clest	0	1	20	29
3	LRT	0	0	0	47
		1	2	3	4*
4	Clest	0	0	1	49
4	LRT	0	0	0	50
		1	2*	3	4
5	Clest	0	44	0	6
5	LRT	0	50	0	0
		1	2*	3	4
6	Clest	0	43	7	0
6	LRT	0	50	0	0
		1	2*	3	4
7	Clest	26	15	6	3
7	LRT	0	46	0	0
		1	2	3*	4
8	Clest	0	16	34	0
8	LRT	0	1	48	1

The true number of groups is denoted by the asterisk.

Here $\mathbf{0}_7$ denotes a seven-dimensional vector of zeros. Any simulation where the Euclidean distance between the two closest observations belonging to different clusters is less than 1 is discarded.

Model 4: $(g = 4, p = 10, n_i = 25$ or 50 with probability 0.5 each), where

$$\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{I}_{10})$$

and where $\boldsymbol{\mu}_i = \boldsymbol{w}_i$ and $\boldsymbol{w}_i$ is a realization of the random variable $\boldsymbol{W}_i$ distributed as

$$\boldsymbol{W}_i \sim N(\mathbf{0}_{10}, 3.6\boldsymbol{I}_{10}).$$

Any simulation where the Euclidean distance between the two closest observations belonging to different clusters is less than 1 is discarded.

Model 5: $(g = 2, p = 3, n_i = 100)$, where

$$\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_{ij}, \boldsymbol{I}_3)$$

and where

$$\boldsymbol{\mu}_{1j} = -0.5 + 0.1(j-1)/99$$

and $\boldsymbol{\mu}_{2j} = \boldsymbol{\mu}_{1j} + 10$.

Model 6: $(g = 2, p = 10, n_i = 100)$, where the simulated observations $\boldsymbol{Y}_{ij} = (\boldsymbol{Y}_{1ij}^T, \boldsymbol{Y}_{2ij}^T)^T$ are formed independently by generating the $\boldsymbol{Y}_{1ij}$ as in Model 5 and by generating the $\boldsymbol{Y}_{2ij}$ as

$$\boldsymbol{Y}_{2ij} \sim N(\boldsymbol{0}_7, \boldsymbol{D}_7)$$

and $\boldsymbol{D}_7$ is a 7×7 diagonal matrix whose vth diagonal element is equal to $(v+3)^2$ $(v = 1, \ldots, 7)$.

Model 7: $(g = 2, p = 10, n_i = 50)$, where

$$\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{I}_{10})$$

and where $\boldsymbol{\mu}_1 = \boldsymbol{0}_{10}$ and $\boldsymbol{\mu}_2 = (2.5, \boldsymbol{0}_9^T)^T$.

Model 8: $(g = 3, p = 13, n_i = 50)$, where

$$\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$$

and where

$$\boldsymbol{\mu}_1 = \boldsymbol{0}, \quad \boldsymbol{\mu}_2 = (2, -2, 2, \boldsymbol{0}_{10}^T)^T \quad \text{and} \quad \boldsymbol{\mu}_3 = (-2, 2, -2, \boldsymbol{0}_{10}^T)^T,$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{O}_{7.3} \\ \boldsymbol{O}_{3.7} & \boldsymbol{I}_{10} \end{pmatrix}$$

and

$$\begin{aligned} (\boldsymbol{\Sigma}_{11})_{uv} &= 1.0 \quad (u = v), \\ &= 0.5 \quad (u \neq v) \end{aligned}$$

for $u, v, = 1, 2, 3$. Here $\boldsymbol{O}_{3.7}$ denotes a 3×7 matrix of zeros.

For each simulated sample, we fitted a g-component normal mixture model, starting with $g = 1$. We kept increasing g until we reached a value of g, g_o, such that the LRT of $H_0 : g = g_o$ versus $H_1 : g = g_o + 1$ was not significant with the P-value assessed by resampling as described in Section 4. The component-covariance matrices were taken to be unrestricted for all but Models 3 and 4 for which they were specified to be equal. For each of the eight simulation models, the value of g_o obtained in this manner on the 50 simulation trials per model are displayed in Table 1.

It can be seen in Table 1 that the relative performance of the LRT with the P-value assessed via resampling is quite encouraging for choosing the number of clusters. The good simulation results for this approach are to be expected since it is favored by having multivariate normal data.

7. Results for microarray data

We further demonstrate the effectiveness of the LRT for the choice of number of clusters by applying it now to some microarray datasets as available in the literature and used in the comparisons of Dudoit and Fridlyand [13].

7.1. Leukaemia data

We firstly consider the clustering of the leukaemia dataset as in [17]. It consists of $M = 72$ tissue samples and $N = 3731$ genes. The 72 samples are made up of 47 cases of acute lymphoblastic leukaemia (ALL) of which there are $n_1 = 38$ ALL B-cell cases and $n_2 = 9$ ALL T-cell cases, along with $n_3 = 25$ cases of acute myeloid leukaemia (AML). We followed the processing steps of Dudoit et al. [13] of (i) thresholding: floor of 100 and ceiling of 16,000; (ii) filtering: exclusion of genes with max/min $\leqslant 5$ and $(\max - \min) \leqslant 500$, where max and min refer, respectively, to the maximum and minimum expression levels of a particular gene across a tissue sample; (iv) the natural logarithm of the expression levels was taken. Before we standardized the genes (the rows of the (logged) data matrix $\boldsymbol{A}$) to have means zero and unit standard deviations over the tissue samples, we first standardized the arrays (the columns of the data matrix $\boldsymbol{A}$) to have zero means and unit standard deviations. This was done in an attempt to remove systematic sources of variation, as discussed, for example, in [14, p. 171].

Obviously, there are far too many genes relative to the tissue samples to fit a normal mixture model directly to the $n = 72$ samples on the basis of all the genes. Thus we used the EMMIX-GENE program [30] to first remove those genes assessed as having little discriminatory capacity across the $n = 72$ tissue samples by fitting a mixture of t distributions to each of the 3731 genes considered separately. This led to 2069 genes being retained where, for the retention of a gene, the threshold for the increase in twice the log likelihood $(-2 \log \lambda)$ was set to be 8 and the minimum cluster size was set to be 5. We then summarized the retained genes by clustering them into $N_o = 40$ groups on the basis of the 72 tissue samples by fitting in equal proportions a mixture of 40 normal components with a common spherical covariance matrix, $\sigma^2 \boldsymbol{I}_{72}$.

An inspection of the heat maps in which the genes within a cluster group are displayed for the 72 tissue samples shows that the second cluster group (containing some 73 genes) is one of the more useful groups for revealing the differences between the tissue samples. We therefore considered the clustering of the 72 tissue samples on the basis of the top 40 genes in the second group of genes.

If we start the iterative fitting of a mixture of $g = 3$ factor analyzers from the external classification of $n_1 = 38$ B-cell ALL cases, $n_2 = 9$ T-cell ALL cases, and $n_3 = 25$ AML cases, we obtain a solution (S_1) of the likelihood equation that leads to an outright clustering C_1 that corresponds almost perfectly with the external classification. However, using 10 random starts and 10 k-means-based starts, we find a nonspurious solution with a higher likelihood value that leads to a different clustering C_2 of the tissues into three clusters. One cluster consists of the 25 AML cases plus the 9 T-cell ALL cases and 2 B-cell ALL cases; the second and third clusters contain 18 each of the remaining 36 B-cell ALL cases. These differences may be of practical interest, but we did not pursue this here as it is outside the scope of this study.

The two-cluster solution has the 25 AML cases plus 12 ALL cases (9 T-cell and 3 B-cell) in one cluster and the remaining 35 B-cell ALL cases in the other. We carried out the LRT test of $g = 2$ versus $g = 3$ component factor analyzers via a resampling approach using $B = 39$ bootstrap samples. As the value of $-2\log\lambda$ for the original sample is greater than the largest of the 39 bootstrap values of $-2\log\lambda$, the P-value is estimated to be less than 0.025. This suggests that there is strong support for $g = 3$ clusters in this set.

7.2. Lymphoma data

The second dataset to be considered here concerns the case study of Alizadeh [2], which measured the gene-expression levels using a specialized cDNA microarray, the Lymphochip. The data consist of $M = 80$ tissue samples and $N = 4062$ genes. The former consist of $n_1 = 29$ cases of B-cell chronic lymphocytic leukaemia (B-CLL), $n_2 = 9$ cases of follicular lymphoma (FL), and $n_3 = 42$ cases of diffuse large B-cell lymphoma (DLBCL). The missing data were imputed as in [13].

We first clustered the 80 tissue samples by fitting a mixture of $g = 3$ factor analyzers to the means of the 40 gene clusters produced by the EMMIX-GENE program [30]. If we start the iterative fitting process from the aforementioned external classification of the tissues, we obtain a solution (S_1) of the likelihood equation that leads to an outright clustering (C_1) that corresponds perfectly with the external classification. However, if we use 10 random and k-means-based partitions to start the iterative fitting, we obtain a nonspurious solution (S_2) at which the likelihood has greater value than for S_1. The clustering C_2 produced by the solution S_2 has one cluster consisting of the 29 B-CLL cases, another consisting of the 9 FL cases and 7 DLBCL cases, and a third cluster consisting of the remaining 35 DLBCL cases. On the basis of the likelihood ratio test, it was concluded that a mixture model of $g = 3$ factor analyzers (with $q = 4$ factors) is adequate for describing the group structure in the dataset ($0.075 < P < 0.1$ using $B = 39$ bootstrap samples).

7.3. Melanoma data

The third dataset we considered concerns the melanoma data of Bittner et al. [8]. It consists of $M = 31$ tissue samples and $N = 3613$ genes. We again used EMMIX-GENE (with the same thresholds as for the leukaemia data) to reduce the number of genes in this set to 571, which were then clustered into 15 groups. As an inspection of the heat maps in which the genes within a cluster group are displayed for the 31 tissue samples shows that the first cluster group (containing some 49 genes) is one of the more useful groups for revealing the separation of the last 19 tissues from the rest, we worked with this cluster of genes for our subsequent testing for cluster structure among the 31 tissues. As noted in [13], there are no a priori classes known for this dataset. The analysis of Bittner et al. [8] suggests that two classes may be present, as they identified a major cluster of 19 tissues. The Clest procedure [13] also yielded two classes, although their clustering has four tissues joined to the 19 member cluster identified in [8].

Application of the likelihood ratio test in conjunction with the fitting of a mixture of g-factor analyzers with $q = 4$ clusters gives a significant result (but close to the borderline) at the 5% level ($0.04 < P < 0.05$, using $B = 99$ bootstrap samples) for the test of $g = 2$ versus $g = 3$. The two-cluster solution has the last 19 tissues along with the 1st, 7th, 8th, and 10th tissues in one cluster with the remaining 9 tissues in the other cluster. If we use the solution obtained from starting with the partition that has the first 12 tissues in one group and the last 19 in another, then we obtain this clustering, but it corresponds to a smaller local maximum.

8. Discussion

One advantage of the mixture model-based approach to cluster analysis is that it provides a sound mathematical basis for clustering and the subsequent testing for group structure in a dataset. A test for the smallest number of components in the mixture model compatible with the data can be formulated in terms of the likelihood ratio statistic. In the present context where the dimension p of the feature vector (the number of genes) is so much greater than the number n of observations (the tissues) to be clustered, the normal mixture model is unable to be fitted directly. This situation is handled by using the EMMIX-GENE program [30], which has the facility for first eliminating those genes with little variation across the tissue samples, and then second, clustering the remaining genes into a manageable number of groups, using essentially Euclidean distance. This assures that highly correlated genes are put in the same cluster. The clustering of the tissue samples can then be undertaken by considering the clusters of genes individually or collectively with each cluster represented by its sample mean. Even with this reduction in the number of genes, the normal mixture model of interest for the clustering of the tissue samples may still not be able to be fitted directly. If this is the case, then we fit mixtures of factor analyzers, whereby the component correlations between the genes are modelled by allowing the distribution of the vector of gene expressions to depend linearly on a small number q of latent (unobservable) variables. It is proposed with this mixture approach that the choice of the number of clusters be made by testing for the smallest number of components in the mixture model compatible with the data. The test can be carried out on the basis of the likelihood ratio test statistic $-2 \log \lambda$ with its null distribution approximated by resampling. In this study, this approach was implemented starting with a single-component factor analyzer and proceeding to add a component factor analyzer into the mixture model until the test for an additional component is nonsignificant. The performance of this approach is demonstrated under simulation models and on microarray cancer datasets as considered previously in the literature. Note that we did not attempt to account for the preprocessing in these applications of the resampling approach.

We did not consider here the related clustering problem of clustering the genes on the basis of the tissue samples. One aim of wishing to identify those genes that have similar gene expressions over the tissue samples might be to find genes under similar regulatory control (assuming that coexpressed genes have similar functional roles). A normal mixture model-based clustering of the genes on the basis of the tissue samples is straightforward in the sense that the normal mixture model can be fitted directly since the number of observations (the genes) is very large relative to the number of feature variables (the tissues). However, for this problem of clustering the genes, there is one condition of a standard cluster analysis that is not satisfied, namely the observations to be clustered (that is, the genes) are not all independent. We can still effect a clustering by proceeding to fit the normal mixture model as if the genes were independently distributed. But tests concerned with the smallest number of components in the mixture model would need to take into account the breakdown in the independence condition [3].

Acknowledgments

We wish to thank the editors and the referees for their valuable comments and suggestions which helped to improve the original version of the paper.

References

[1] M. Aitkin, D. Anderson, J.P. Hinde, Statistical modelling of data on teaching styles (with discussion), J. Roy. Statist. Soc. Ser. B 144 (1981) 419–461.

[2] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Truc, X. Yu, J.I. Powell, L. Yang, et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.

[3] D.B. Allison, G.L. Gadbury, M.S. Heo, J.R. Fernández, C.-K. Lee, T.A. Prolla, R. Weindruch, A mixture model approach for the analysis of microarray gene expression data, Comput. Statist. Data Anal. 39 (2002) 1–20.

[4] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Nat. Acad. Sci. 96 (1999) 6745–6750.

[5] G.A. Barnard, Contribution to the discussion of paper by M.S. Bartlett, J. Roy. Statist. Soc. Ser. B 25 (1963) 294.

[6] A. Ben-Dor, R. Shamir, Z. Yakhini, Clustering gene expression patterns, J. Comput. Biol. 6 (1999) 281–297.

[7] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, et al., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinaom subclasses, Proc. Nat. Acad. Sci. 98 (2001) 13790–13795.

[8] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, et al., Molecular classification of cutaneous malignant, Nature 406 (2000) 536–540.

[9] J. Bryan, K.S. Pollard, M.J. van der Laan, Paired and unpaired comparison and clustering with gene expression data, Statist. Sinica 12 (2002) 87–110.

[10] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Comm. Statist. Theory Methods 3 (1974) 1–27.

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

[12] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. Roy. Statist. Soc. Ser. B 39 (1977) 1–38.

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

[14] S. Dudoit, J. Fridlyand, Classification in microarray experiments, in: T. Speed (Ed.), Statistical Analysis of Gene Expression Microarray Data, Chapman & Hall/CRC, Boca Raton, FL, 2003, pp. 93–158.

[15] M.B. Eisen, P.T. Spellmann, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Nat. Acad. Sci. 95 (1998) 14863–14868.

[16] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Amer. Statist. Assoc. 78 (1983) 553–584.

[17] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeck, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

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

[19] J.A. Hartigan, Statistical theory in clustering, J. Classification 2 (1985) 63–76.

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

[21] C. Hennig, Breakdown points of maximum likelihood-estimators of location-scale mixtures, 2002, Unpublished manuscript.

[22] D.C. Hoaglin, Using quantiles to study shape, in: F. Mosteller, J.W. Tukey (Eds.), Explaining Data Tables, Trends, and Shapes, Wiley, New York, 1985, pp. 417–460.

[23] A.C.A. Hope, A simplified Monte Carlo significance test procedure, J. Roy. Statist. Soc. Ser. A 30 (1968) 582–598.

[24] W. Huber, A. von Heydebreck, H. Sueltmann, A. Poustka, M. Vingron, Parameter estimation for the calibration and variance stabilization of microarray data, Statist. Appl. Genetics Molecular Biol. 2 (1) (2003) Article 3.

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

[26] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data, Wiley, New York, 1990.

[27] W.J. Krzanowski, Y. Lai, A criterion for determining the number of groups in a dataset using sum of squares clustering, Biometrics 44 (1985) 23–34.

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

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

[30] G.J. McLachlan, R.W. Bean, D. Peel, A mixture model-based approach to the clustering of microarray expression data, Bioinformatics 18 (2002) 413–422.

[31] G.J. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley, New York, 1997.

[32] G.J. McLachlan, D. Peel, On a resampling approach to choosing the number of components in normal mixture models, in: L. Billard, N.I. Fisher (Eds.), Computing Science and Statistics, Vol. 28, Interface Foundation of North America, Fairfax Station, VA, 1997, pp. 260–266.

[33] G.J. McLachlan, D. Peel, Robust cluster analysis via mixtures of multivariate *t*-distributions, in: A. Amin, D. Dori, P. Pudil, H. Freeman (Eds.), Lecture Notes in Computer Science, Vol. 1451, Springer, Berlin, 1998, pp. 658–666.

[34] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.

[35] G.J. McLachlan, D. Peel, Mixtures of factor analyzers, in: P. Langley (Ed.), Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 2000, pp. 599–606.

[36] G.J. McLachlan, D. Peel, K.E. Basford, P. Adams, Fitting of mixtures of normal and *t*-components, J. Statist. Software 4 (1999) 2.

[37] G.J. McLachlan, D. Peel, R.W. Bean, Modelling high-dimensional data by mixtures of factor analyzers, Comput. Statist. Data Anal. 41 (2003) 379–388.

[38] G. Parmigiani, E.S. Garrett, R.A. Irizarry, S.L. Zeger (Eds.), The Analysis of Gene Expression Data, Springer, New York, 2003.

[39] D. Peel, G.J. McLachlan, Robust mixture modelling using the *t*-distribution, Statist. Comput. 10 (2000) 335–344.

[40] K.S. Pollard, M.J. van der Laan, Statistical inference for simultaneous clustering of gene expression data, Math. Biosci. 176 (2002) 99–121.

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

[42] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.

[43] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C.T. Agular, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, et al., Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Natur. Med. 8 (2002) 68–74.

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

[45] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. Roy. Statist. Soc. Ser. B 63 (2001) 411–423.

[46] E.P. Xing, R.M. Karp, CLIFF: clustering of high-dimensional MICROARray data via iterative feature filtering using normalized cuts, Bioinformatics 17 (2001) S306–S315.

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