Classification of microarray data with factor mixture models

Francesca Martella
Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università degli Studi di Roma “La Sapienza”, P.le A. Moro, 5-00185, Rome, Italy

ABSTRACT
Motivation: The classification of few tissue samples on a very large number of genes represents a non-standard problem in statistics but a usual one in microarray expression data analysis. In fact, the dimension of the feature space (the number of genes) is typically much greater than the number of tissues. We consider high-density oligonucleotide microarray data, while the expression level is associated to an ‘absolute call’, which represents a qualitative indication of whether or not a transcript is detected within a sample. The ‘absolute call’ is generally not taken in consideration in analyses.

Results: In contrast to frequently used cluster analysis methods to analyze gene expression data, we consider a problem of classification of tissues and of the variables selection. We adopted methodologies formulated by Ghahramani and Hinton and Rocci and Vichi for simultaneous dimensional reduction of genes and classification of tissues; trying to identify genes (denominated ‘markers’) that are able to distinguish between two known different classes of tissue samples. In this respect, we propose a generalization of the approach proposed by McLachlan et al. by advising to estimate the distribution of log LR statistic for testing one versus two component hypothesis in the mixture model for each gene considered individually, using a parametric bootstrap approach. We compare conditional (on ‘absolute call’) and unconditional analyses performed on dataset described in Golub et al. We show that the proposed techniques improve the results of classification of tissue samples with respect to known results on the same benchmark dataset.

Availability: The software of Ghahramani and Hinton is written in Matlab and available in ‘Mixture of Factor Analyzers’ on http://www.gatsby.ucl.ac.uk/~zoubin/software.html while the software of Rocci and Vichi is available upon request from the authors.

Contact: francesca.martella@uniroma1.it

1 INTRODUCTION
In general, there are two different types of DNA microarrays: spotted microarrays and oligonucleotide microarrays. There are several important differences between these two types of microarrays. While the former are obtained using special spotting robots, the latter are synthesized, often using photolithographic technologies. The wealth of gene expressions data, which is becoming available, poses numerous statistical problems ranging from the analysis of images produced by microarray experiments to biological interpretation of analysis output. Here, we focus on study of leukemias using gene expression data. By allowing the monitoring of expression levels for thousands of genes simultaneously, such techniques may lead to a more complete understanding of the molecular variations among leukemias and hence to finer information about how genes work in these diseases. There are three main types of statistical problems associated with leukemia studies:

1) the identification of new/novel leukemia classes using gene expression profiles (cluster analysis or unsupervised learning/class discovery/ clustering);
2) the classification of malignancies into known classes (discriminant analysis/supervised learning/class prediction);
3) the identification of ‘marker’ genes characterizing the different classes (variables selection).

In the biostatistical literature, various analysis procedures have been applied to microarray data: in particular hierarchical clustering methods (Eisen et al., 1998; Alizadeh et al., 2000; Bittner et al., 2000; Tibshirani et al., 1999), self-organizing maps (SOM) (Tamayo et al., 1999; Kohonen, 1999; Golub et al., 1999), single linkage k-method (Li et al., 2001a, b), support vector machines (Cortes and Vapnik, 1995) and discriminant analysis methods (Vandeginste et al., 1998) have played a major role. Effective results in applications have been reported for many analysis approaches, but no single method has emerged as the better in the gene expression analysis. In particular, concerning the unsupervised learning, most of the proposed clustering algorithms are heuristically motivated, and the issues of determining the correct number of clusters and of choosing a good clustering algorithm are not yet rigorously solved.

An important class of clustering techniques that can provide alternative solutions to these problems is that of clustering algorithms based on probability models. In general, those models work well both in an unsupervised and a supervised context. In particular, model-based approach assumes that the data are generated by a finite mixture of underlying probability distributions (each one corresponding to a cluster) such as multivariate Normal distributions. With model-based approach, the problems of determining the number of clusters (in an unsupervised learning) and of choosing an appropriate clustering method become a problem of model selection (e.g. Dasgupta and Raftery, 1998; Celeux and Govaert, 1993; McLachlan et al., 2002). A model-based clustering has the advantage to define a probabilistic framework that allows to select the number of clusters in the data according to well-known and used selection criteria such as Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). In other words, we choose the number of clusters and, therefore, the best clustering results using the log-likelihood function penalized by the number of parameters.
of the mixture model. Other famous selection criteria are the Approximate Weight of Evidence (AWE, see Banfield and Raftery, 1993) and the model selection approach based on cross-validated likelihood proposed by Smyth (2000). For a complete review of model selection criteria see McLachlan and Peel (2000).

However, if the issue is to fit a finite mixture model in order to cluster gene expression experiments, the conceptual and computational limitations make the use of standard algorithm for maximum likelihood estimation rather complicated. In fact, observations are usually considered as $n$ independent realizations of $p$-dimensional random variables, where the sample dimension is greater than the dimension of each observation, and the latter in turn is not greater than dimensions of cluster to be formed, in order to avoid singular estimates of covariance matrix corresponding to each cluster. Therefore, if the aim is to classify few tissue samples characterized by a very large number of genes, it is clear that we can encounter some problems with parameter estimation and we consequently need to proceed to dimensional reduction before using a clustering algorithm.

Often principal component analysis (PCA) has been applied to reduce the number of the variables, to apply a clustering algorithm for grouping units into homogeneous classes on the basis of principal components. The hope for using PCA prior to cluster analysis is that PCs may ‘extract’ the essential information about the cluster structure in dataset. Since PCs are uncorrelated and ordered, the first few PCs, which contain most of the variations in the data, are usually used in cluster analysis (Jolliffe et al., 1980). This procedure, named ‘tandem analysis’ by Arabie and Hubert (1994), provides PCs and an optimal classification, minimizing two different target functions that may work in opposite directions. That is because PCA may identify PCs that do not contribute much to perceive the clustering structure in data but, on the contrary, may obscure or mask it.

To overcome this problem, several authors proposed methodologies for simultaneous classification and reduction of the analyzed data.

In this paper, we performed simultaneous dimensional reduction of genes and classification of tissues comparing the results of two analyses: the first analysis considers the information provided by the ‘absolute call’ (conditional analysis) while the other one does not consider the call (complete analysis). The adopted statistical techniques have been formulated by Ghahramani and Hinton (1996) and Rocci and Vichi (2002).

The gene expression levels analyzed in this work have been taken using high-density oligonucleotide microarrays data produced by Affymetrix TM corporation (dataset described in Golub et al., 1999). According to this procedure, each gene has been assigned a measure of expression level, and an ‘absolute call’, which represents a qualitative indication determining whether there is sufficient mRNA present on certain spots of the array and consists of three values, Absent [A], Marginal [M] and Present[P]. An Absent absolute call indicates that there is insufficient mRNA concentration presence on the spot, whereas a Present absolute call indicates otherwise. A Marginal absolute call means that there is not enough information to make the call Absent or Present. However, it has been noted that Affymetrix software is not much sensitive to low expression values, and then it may assess an ‘Absent absolute call’ even if the transcription has correctly happened. In literature, the ‘absolute call’ is not generally taken into account. Aris and Recce (2002) have focused on genes that are selectively expressed rather than on differential expression levels.

Furthermore, we compared the performance of these techniques with other analyses of the same dataset, which have been previously published in literature. We aimed at identifying genes (denominated ‘markers’) that are able to discriminate between classes of tissue samples through the imposition of a threshold on the log likelihood ratio statistic for testing one versus two component hypothesis in the mixture model, according to McLachlan et al. (2002).

We have evaluated the obtained results fitting Ghahramani and Hinton (1996) model only on ‘marker’ genes.

The paper is organized as follows. Section 2 contains the dataset description with pre-processing and normalization steps. Section 3 discusses the adopted statistical techniques while the corresponding results are presented in Section 4. Finally, Section 5 summarizes our findings.

2 MICROARRAY DATA

Gene expression microarrays typically contain thousands of oligonucleotides or cDNAs, which correspond to transcripts of many different genes.

DNA microarrays typically correspond to two main types of technical platforms: one is based on small glass slides where cDNAs or oligonucleotides (70–80mers) have been spotted. The other, which is more expensive, is based on photolithographic techniques to synthesize 25mer oligonucleotides on a silicon wafer (Affymetrix Inc.). These platforms differ in their technology for labeling RNA molecules, but the main difference concerns the methods to assess the gene transcript levels. Gene expression ratios for each spot corresponding to ‘target’ and ‘control’ sample for glass slides absolute gene and expression levels for each spot in the case of Affymetrix microarrays. Each technology has its advantages and disadvantages, and the appropriate choice typically depends on the specific experiment to be prosecuted.

Here, we will illustrate only the Affymetrix microarray procedure, since the analyzed data come from this type of technology.

Affymetrix microarrays generate a gene expression profile of one sample; each gene expression is represented by (typically 20) 25mer oligonucleotides. In addition to these perfect-match oligonucleotides, each 25mer comes with a negative control oligonucleotide that contains a mismatch at central position. The integration of the expression levels for each of the 20 perfect-match-mismatch oligonucleotide sets generates a value for the expression of a particular gene. mRNA is back-transcribed into cDNA, which is used in vitro reaction to generate biotinylated cRNA. After fragmentation, this cRNA is hybridized to microarrays, washed and stained with PE-conjugated streptavidin, and subsequently scanned on a laser scanner. The final output consists of two matrices with rows corresponding to genes and columns corresponding to tissue samples replicated twice. The first of these contains the absolute gene expression profiles while the second contains the corresponding ‘absolute call’.

2.1 Data description

The analyzed dataset, described in Golub et al. (1999) comes from a study of gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix
high-density oligonucleotide arrays (HU6800chip). The intensity values have been re-scaled such that overall intensities for each array are equivalent (see http://www.broad.mit.edu/mpr/publications/projects/Leukemia/protocol.html).

The dataset contains 7129 human genes taken from 72 tissue samples: 47 samples of acute lymphoblastic leukemia, ALL (38 hit in intermediate precursors of ‘lymphocytes’ B, ‘B-cell’, and 9 hit in intermediate precursors of ‘lymphocytes’ T, ‘T-cell’) and 25 samples of acute myeloid leukemia, AML. The 72 tissue samples have been divided in two sets: a training set containing 38 tissue samples, 27 ALL (8 T-cell and 19 B-cell) and 11 AML; and a test set of 34 tissue samples, 20 ALL (1 T-cell and 19 B-cell) and 14 AML.

In the training set, used in order to estimate a discriminant function for the two groups (ALL and AML), only bone marrow samples are present, while the test set, used to evaluate the obtained separation, contains also peripheral blood samples carrying further variability in the analysis.

Each sample has been associated to gene expression level value and corresponding ‘absolute call’ (Present [P], Absent [A], Marginal [M]). The absolute calls are generated by Affymetrix scanning software and are an indication of the confidence in measured expression values. In other words, the back-transcribed gene, made during the drawn of RNA, can be actually Present/Absent/ Marginal. However, it has been noted that Affymetrix software is not sensitive to low gene expression values, and then it may assess an ‘Absent call’ even if the gene transcription has correctly occurred. In analyses of this dataset, Golub et al. (1999) and McLachlan et al. (2002) ignored the absolute call information, basing their analysis only on gene expression levels. Airis and Recce (2002) have focused on genes that are selectively expressed rather than on differential expression levels.

### 2.2 Pre-processing and normalization

After generation of the gene expression profiles, it is necessary to adopt some transformation on these measurements in order to reduce chip effects, background intensity, variations from RNA extraction, labeling, dye efficiency and other variability sources related to experiment. Several normalization and transformation methods have been proposed (see e.g. Alizadeh et al., 2000; Dopazo et al., 2001; Bolstad et al., 2003; Yeung et al., 2001a, b), but no ‘standard’ method exists. We have applied a pre-processing scheme which resembles the one followed by McLachlan et al. (2002) and Dudoit et al. (2002) and recommended by Pablo Tamayo, one of the authors of Golub et al. (1999). We used this scheme in order to allow for direct comparability with previous analyses on the same dataset. The steps were as follows:

1. Restrict gene expression values to the range (100, 16 000). That is, gene expression levels above 16 000 will be cut to 16 000, since fluorescence saturation is present above this level and corresponding values cannot be reliably measured. Furthermore, gene expression levels below 100 will be set to 100 since they play the same role of values equal to 100 in the analysis.

2. Exclusion of genes with (max/ min) ≤ 5 or (max – min) ≤ 500, where max and min refer to the maximum and minimum intensities for a particular gene across mRNA samples, respectively, i.e. exclusion of genes with reduced variability across samples being analyzed.

3. Use base 10 logarithmic transformation.

4. Standardize each column to have zero mean and unit standard deviation and then standardize each row to have zero mean and unit deviation. This was done to remove systematic sources of variation, as discussed in Dudoit and Fridlyand (2003).

### 3 METHODOLOGIES

Before we proceed to describe the methodologies we adopted for the present analysis, let us start summarizing the mixture model approach.

Suppose data consist of $n$ independent $p$-dimensional observations $y_1, \ldots, y_n$. A mixture model approach assumes that each observation $y_i (i = 1, \ldots, n)$ is drawn from a mixture of a fixed number $K$ of groups (corresponding to mixture component densities) in some unknown proportions $\pi_1, \ldots, \pi_K$. In other words, $y_i$ is a realization of a random vector $Y_i$ corresponding to $p$ features measured on the $i$-th statistical unit, with density function defined by

$$ f(y_i; \phi) = \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k), $$

where $f_k(y_i; \theta_k)$ denotes the $k$-th component density with parameter vector $\theta_k$, the $\pi_k$ represent mixing weights with $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$, $k = 1, \ldots, K$ and $\phi = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$.

Specifying for $k$-th component a $p$-variate Normal density function $\varphi$ with mean vector $\mu_k$ and covariance matrix $\Sigma_k$, we have

$$ f(y_i; \phi) = \sum_{k=1}^{K} \pi_k \varphi(y_i; \mu_k, \Sigma_k) $$

$$ = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp \left\{ -\frac{1}{2} (y_i - \mu_k)^\top \Sigma_k^{-1} (y_i - \mu_k) \right\}. $$

A sample of indicator vectors or labels $z = (z_1, \ldots, z_n)$, with $z_i = (z_{i1}, \ldots, z_{iK})$, $z_{ik} = 1$ or 0 according to the fact that $y_i$ is arising from $k$-th mixture component or not, is associated to the observed data $y$. The sample $z$ can be known when we are in a discriminant analysis context, where the problem is essentially to predict an indicator vector $z_{n+1}$ from a new observed data vector $y_{n+1}$. If the sample is unknown, we are in a cluster analysis context, i.e. the estimation of the $z_{ik}$ are of primary interest. In both cases, the vector of unknown parameters is represented by

$$ \phi = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K). $$

An estimate of $\phi$ can be derived from the maximum likelihood estimate $\hat{\phi}$ of the mixture parameters obtained, for instance, by the EM algorithm, by assigning each $y_i$ to the component $k$ providing the largest conditional probability, $w_{ik} (k = 1, \ldots, K)$, that $y_i$ arises from it, using Maximum a Posterior (MAP) approach. The MAP approach is as follows:

$$ z_{ik} = \begin{cases} 1 & \text{if } k = \arg\max_{k=1,\ldots,K} w_{ik} \vspace{0.2cm} \\ 0 & \text{otherwise} \end{cases}. $$

(3)
where
\[ w_{ik} = \frac{\pi_k \varphi(y_{ik}; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \varphi(y_{ij}; \mu_j, \Sigma_j)} \] (4)

3.1 Ghahramani and Hinton model (1996)

Ghahramani and Hinton, 1996 proposed a model that concurrently performs clustering and, within each cluster, local dimensional reduction. In their factor analysis model, the \( p \)-dimensional data vector \( y_i \) is modeled as
\[ y_i = \mu + Bu_i + e_i, \] (5)
where \( \mu \) is the mean vector, \( B \) is a \( p \times q \) matrix of factor loadings, \( u_i \) is a \( q \)-dimensional \((q < p)\) vector of latent or unobservable variables known as factors, which are assumed to be i.i.d. as \( N(0, I_q) \), where \( I_q \) denotes the \( q \times q \) identity matrix. Furthermore, \( e_i \) are i.i.d. with Normal distributions \( N(0, D) \), \( D = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2) \), which are assumed to be independent of \( u_i \).

According to this model, the distribution of \( y_i \) is given by
\[ y_i \sim N(\mu, BB' + D). \] (6)

The goal of factor analysis is to find \( B \) and \( D \) that best model the covariance structure of \( y_i \).

Ghahramani and Hinton, 1996 considered a mixture of \( K \) factor models (5), called mixture of factor analysers (\( m \)-FA), each factor analyzer depends on a set of \( q \) latent factors through a component-specific factor loading matrix \( B_k \) and on different mean vectors \( \mu_k \). Thus, each \( y_i \) is a mixture of \( K \) Normal component densities with proportions \( \pi_1, \ldots, \pi_K \), that is
\[ f(y_i; \phi) = \sum_{k=1}^{K} \pi_k \varphi(y_i; \mu_k, \Sigma_k), \] (7)
where \( \Sigma_k = B_kB_k' + D_k \), \( k = 1, \ldots, K \).

The parameters of this model are
\[ \{(\mu_k, B_k)_{k=1}^K, \pi_1, \ldots, \pi_{K-1}, D\}. \]

Here, we do not explain the details of their model, but it is important to note that the mixture of factor analyzers is, in essence, a dimensional reduction of a mixture of Normal distributions, where each factor model fits a Normal distribution to a portion of data, with weights given by posterior probabilities, \( w_{ik} \) \( k = 1, \ldots, K, i = 1, \ldots, n \). Since the covariance matrix for each distribution is specified through the lower dimensional factor loading matrices, the model has \( pq - (q^2 - q)/2 \) \(+ p \) rather than \( Kp(p + 1)/2 \) parameters to modeling covariance structure. It means that if \( q \) is sufficiently smaller than \( p \), \( \Sigma_k = B_kB_k' + D_k \) imposes stronger restrictions on covariance matrices reducing the free parameters number.

3.2 Rocci and Vichi model (2002)

Rocci and Vichi, 2002 proposed a classification model that has some links with the factor analysis model proposed by Ghahramani and Hinton, 1996. More precisely, they proposed a two-way model for simultaneous dimensional reduction and classification assuming that the observed data come from a finite mixture of multivariate Normal distributions. Constraining the mean vectors of each mixture component density to lie in a reduced subspace, model parameters are estimated through maximum likelihood using an Expectation Conditional Maximization (ECM) algorithm which is proposed by the authors.

They consider the mixture model (2) and represent the mean vector \( \mu_k = [\mu_{k1}, \ldots, \mu_{kp}] \) of each component density, as a function of \( q < (p) \) latent variables according to the following bilinear model:
\[ \mu_k = Bu_k, \] (8)
where \( B \) is a factor-loading matrix \( [p \times q] \), while \( u_k = [u_{k1}, \ldots, u_{kq}] \) represents the factor score vector of the \( k \)-th component.

Let \( y_1, \ldots, y_n \) be \( n \) i.i.d. observations, the maximized log-likelihood of mixture model (1) can generally be decomposed as (Hathaway, 1986):
\[ l(\phi) = \sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_k \varphi(y_i; \theta_k) \right) 
= \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} \log (\pi_k \varphi(y_i; \theta_k)) - \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} \log (w_{ik}), \] (9)
\( \phi \) denoting the maximum likelihood estimate of the corresponding parameter vector, \( w_{ik} \) represents the posterior probability that the \( i \)-th unit belongs to the \( k \)-th component of the mixture subject to constraint \( \sum_{k=1}^{K} w_{ik} = 1 \) \( i = 1, \ldots, n \). The algorithm can be formulated by maximizing the log-likelihood function with respect to only a subset of model parameters conditionally upon the others.

For further details on parameter estimation and the ECM algorithm see Rocci and Vichi (2002).

The main difference between Ghahramani and Hinton (1996) and Rocci and Vichi (2002) models is that the latter assumes the mean vectors of the Normal mixture components are constrained to lie in a common reduced subspace to identify the latent factors that best explain the variability between groups (i.e. the variables with the greatest discriminant power between groups). On the other hand, the former approach assumes unconstrained component-specific mean vector and controls the number of parameters by modeling the specific-covariance structures, \( \Sigma_k = B_kB_k' + D_k \) \( k = 1, \ldots, K \) and, finally, explains the correlations between genes through factors, \( u \).

This provides an intermediate model between the independence and unrestricted model able to capture some interesting correlation structure in the data, but also without fitting a full covariance matrix.

It is interesting to note that the number of parameters to be estimated in the model of Ghahramani and Hinton (1996) is
\[ [pq - (q^2 - q)/2]K + pK + K - 1 \]

Table 1. Gene number used in different analyses

<table>
<thead>
<tr>
<th>Method</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dudoit et al. (2002)</td>
<td>6871</td>
<td>3571</td>
<td>38</td>
<td>34</td>
</tr>
<tr>
<td>Golub et al. (1999)</td>
<td>7129</td>
<td>1000</td>
<td>38</td>
<td>34</td>
</tr>
<tr>
<td>Complete analysis</td>
<td>7129</td>
<td>3571</td>
<td>38</td>
<td>34</td>
</tr>
<tr>
<td>Conditional analysis</td>
<td>350</td>
<td>242</td>
<td>38</td>
<td>34</td>
</tr>
</tbody>
</table>
### Table 2. Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Genes</th>
<th>Clusters</th>
<th>Dimensions</th>
<th>Training errors</th>
<th>Test errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dudoit et al. (2002)</td>
<td>3571</td>
<td>2</td>
<td>—</td>
<td>1.38</td>
<td>2.34</td>
</tr>
<tr>
<td>McLachlan et al. (2002)</td>
<td>2015</td>
<td>2</td>
<td>8</td>
<td>1.72</td>
<td>—</td>
</tr>
<tr>
<td>Golub et al. (1999)</td>
<td>1000</td>
<td>2</td>
<td>—</td>
<td>2.38</td>
<td>5.34</td>
</tr>
<tr>
<td>Ghahramani and Hinton (1996)</td>
<td>3571</td>
<td>2</td>
<td>2</td>
<td>—</td>
<td>2.34</td>
</tr>
<tr>
<td>Rocci and Vichi (2002)</td>
<td>242</td>
<td>2</td>
<td>2</td>
<td>—</td>
<td>2.34</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>2</td>
<td>2</td>
<td>—</td>
<td>2.34</td>
</tr>
<tr>
<td></td>
<td>3571</td>
<td>2</td>
<td>2</td>
<td>—</td>
<td>2.34</td>
</tr>
</tbody>
</table>

### Table 3. Number of parameters to be estimated in the analyses

<table>
<thead>
<tr>
<th>Method</th>
<th>Genes</th>
<th>Clusters</th>
<th>Dimensions</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghahramani and Hinton (1996)</td>
<td>3571</td>
<td>2</td>
<td>2</td>
<td>17856</td>
</tr>
<tr>
<td></td>
<td>242</td>
<td>2</td>
<td>2</td>
<td>1211</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>2</td>
<td>2</td>
<td>221</td>
</tr>
<tr>
<td>Rocci and Vichi (2002)</td>
<td>3571</td>
<td>2</td>
<td>2</td>
<td>10718</td>
</tr>
<tr>
<td></td>
<td>242</td>
<td>2</td>
<td>2</td>
<td>731</td>
</tr>
</tbody>
</table>

while in Rocci and Vichi (2002)

\[
pq + Kq - q^2 + K - 1 + [Kp(p + 1)].
\]

In particular, by assuming a common and spherical covariance matrix \( \Sigma \) in the model of Rocci and Vichi (2002) and a covariance matrix \( \Sigma = BB' + D \) in the model of Ghahramani and Hinton (1996) the number of parameters of Rocci and Vichi (2002) is considerably lower than the number of parameters of Ghahramani and Hinton (1996) (see Table 3, where the exact number of parameters estimated during the applications is summarized). In other words, if it is possible to assume a common and spherical covariance matrix for the density function of the considered data and if the results obtained of both models are similar, then the model of Rocci and Vichi (2002) is more convenient in terms of statistical and computational complexity.

### 3.3 ‘Marker’ gene selection

To select those genes, called ‘marker’ genes, which more effectively discriminate between tissues belonging to the two classes (e.g. healthy and unhealthy samples), McLachlan et al. (2002) suggest to use a formal test based on the log likelihood ratio statistic \(-2 \log(\lambda)\) to test the null hypothesis of one component versus the alternative of two components in the finite mixture. Their idea is to fit one and two component mixture models for each gene considered individually, and to take it to be as ‘marker’ if the log likelihood ratio statistic is greater of a specified threshold.

Following this idea, we have estimated the distribution of \(NLR\) statistic \(-2 \log(\Lambda)\) using a parametric bootstrap approach, and obtained the threshold as the 95th percentile of the estimated distribution. This approach, introduced by McLachlan (1987) in order to determine the number of components in a finite mixture model, has been successively discussed and motivated by Feng and McCullagh (1996). The idea is to use the bootstrap replications to provide a test of approximate size \(\alpha\). That test rejects the null hypothesis if \(-2 \log(\lambda)\) is greater than the \(j\)-th order statistic of the \(B\) bootstrap replications, which can be taken as an estimate of the quantile of order \(j/(B + 1)\). We have followed the steps:

1. Let \(y_i\) be the \(j\)-gene expression level;
2. Draw of an \(n\)-dimensional sample, \(y_i'\), from a Normal distribution with mean and variance equal to the mean and the variance of the gene \(y_i\) selected in the point 1.
3. Fit \(K = 1\) and \(K = 2\) component mixture models to \(y_i'\);
4. Calculate the test statistic
   \[-2 \log(\lambda) = - 2 \log \left( \frac{L_1}{L_2} \right)\]
   for gene \(j\), where \(L_1\) is the maximized likelihood corresponding to \(K = 1\) component mixture model, while \(L_2\) is the maximized likelihood corresponding to \(K = 2\) component mixture model.

This process is repeated independently a number of times, \(B\), and the values of \(-2 \log(\lambda)\) evaluated from the successive bootstrap samples are used to assess the distribution \(-2 \log(\lambda)\) under the null hypothesis. Finally, we have taken the 95th percentile of this distribution as the threshold for the test statistic.

McLachlan and Peel (1997) underlined that the replicate bootstrap number, \(B\), must be rather high in order to reduce the sensitivity of the EM algorithm to the initial values choice; in fact, even if this test may have approximatively required size \(\alpha = 0.05\) for low values of \(B\), the corresponding power may be well below its limit value, obtained with \(B \rightarrow \infty\). For this reason we set \(B = 1000\), a value which can be considered adequate to the task.

### 4 RESULTS

First, we have used the whole set of gene expression data (complete analysis) summarizing it in a \(72 \times 7129\) matrix \(Y = (y_{ip})\), where \(y_{ip}\) denotes the base 10 logarithm of the expression level for gene \(p\) in tumor mRNA sample \(i\). Then, we have only focused on the 350 genes whose transcription was ‘Present’ or ‘Marginal’ for each sample (conditional analysis) summarizing the data by a \(72 \times 350\) matrix \(Y' = (y'_{ip})\), where \(y'_{ip}\) denotes the expression level for gene \(p'\) in the \(i\)-th mRNA sample.

In this latter case, we have lost some information, since the Affymetrix software is able to target as ‘Present’ only genes with an high expression level; therefore we have excluded also genes characterized by a low expression level, which may be meaningful for the analysis. On the other hand, we assumed that ‘Present’ genes will give some additional interesting biological information.

With respect to the complete analysis, after the pre-processing described in Section 2.2, we have selected 3571 genes. On the basis of these genes, we have fitted a mixture of \(K = 2\) Normal factor analyzers with a number of factors \(q = 2\) and with a common and spherical covariance matrix \(\Sigma\) according to the model by Rocci and Vichi (2002) and a mixture of \(K = 2\) Normal factor analyzers with a number of factors \(q = 2\) and with covariance matrix \(\Sigma = BB' + D\) according to the model by Ghahramani and Hinton (1996) with the
aim to classify the 38 tissues of the training set into a subspace with dimension 2. Both models have correctly separated the 38 samples in two groups. Using the parameter estimates obtained on the training set, we have classified the 34 tissues of the test set; also in this case, we have used posterior probabilities to allocate units to the two group. We have observed two misclassifications for both methods.

We have gone on as above for the conditional analysis, but considering only 242 genes (selected by 350 genes, also in this case, using the pre-processing described in the Section 2.2). We have obtained the same results of the complete analysis; even if, in the complete analysis the posterior probabilities of misclassifications are higher than in the conditional analysis because we have considered more genes and so we have a lower stability of model.

In particular, concerning the conditional analysis, the AML sample no. 54, is correctly classified by Rocci and Vichi (2002) method, while using the Ghahramani and Hinton (1996) method it is incorrectly classified into the ALL cluster. It must be noted, however, that this classification has been done with a posterior probability equal to 0.51 with respect to ALL cluster and 0.49 with respect to AML cluster. Moreover, the AML sample no. 54 is incorrectly classified in Golub et al. (1999) analysis, but also in this case, being the posterior probability near to 0.50, it can be considered as a result of the difficulties in separating the classes rather than a ‘real’ misclassification.

For the ALL sample no. 71, we have obtained a similar result. It is incorrectly classified if we used the Rocci and Vichi (2002) method, while using the Ghahramani and Hinton (1996) method it is correctly classified into the ALL cluster; also in this case, posterior probabilities are equal to, respectively, 0.52 for AML cluster and 0.48 for ALL cluster.

For both models, we observed a strong misclassification result: the AML sample no. 66. In fact, it has a posterior probability equal to 0.80 for ALL cluster and 0.20 for AML cluster. The AML sample no. 66 results incorrectly classified also into different analyses (e.g. Golub et al., 1999 and Dudoit et al., 2002), so we can conclude that it is an outlying sample since its expression gene levels have a behavior more similar to that of a generic ALL sample rather than that of a AML sample.

We have used the smallest set of genes (242) in order to deal with the ‘marker’ gene selection problem. As it is described in Section 3.3, we have considered a gene as ‘marker’ if the log likelihood ratio statistic for testing one versus two component hypothesis in the mixture model for that gene considered individually is greater of a specified threshold. The threshold is obtained as 95th percentile of the assessed distribution of $-2\log(\lambda)$ evaluated using a parametric bootstrap approach. Thus, we have selected 44 ‘marker’ genes. On the basis of these genes, we have fitted mixtures of $K = 2$ Normal factor analysers with a number of factors equal to $q = 2$ according to the methodology proposed by Ghahramani and Hinton (1996) to classify the 38 tissues of the training set into a subspace with reduced dimension. We have evaluated this classification on the 34 tissues of the test set, as below and obtained the similar results, which are however more robust with respect to the dependence from initial values, than those obtained through two previous analyses.

5 DISCUSSION

We have showed that the models formulated by Ghahramani and Hinton (1996) and Rocci and Vichi (2002), which simultaneously perform clustering and dimensional reduction using mixture model, have improved the results of classification of the 72 tissue samples described in Golub et al. (1999) with respect to known results on the same benchmark dataset.

In particular, we have formulated a ‘marker’ gene selection model which, in conjunction with the use of the ‘absolute call’, has allowed to select only 44 from 7129 initial genes having a low loss of information with respect to the quality of classification of tissues belonging to the test set.

Finally, we have verified that fitting the same model using only 44 genes, selected through a parametric bootstrap procedure, we obtained similar results. So, we can conclude that the selected genes can be really considered as being able to discriminate between AML and ALL tissues.

An issue that is the subject of future exploration is to investigate the biological significance of clustering and dimensional reduction results. We also plan to investigate a ‘good’ way of evaluating the correct number of clusters in the absence of initial information on classes. Moreover, an interesting future task would be to deal with the block clustering methods in a mixture approach to allow simultaneous clustering of objects (samples) and variables (genes). Finally, a future research would be to mix the two models by reducing the mean vector and the covariance matrix.

Conflict of Interest: none declared.

REFERENCES


Arabie,Ph. and Hubert,L. (1994) Iterative Projection Strategies for the Least-squares Fitting of Graph Theoretic Structures to Proximity Data, Research Report RR-94-02, Department of Data Theory, University of Leiden; 62.


