Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

The full Bayesian significance test for mixture models: results in gene expression clustering

M.S. Lauretto, C.A.B. Pereira and J.M. Stern

Instituto de Matemática e Estatística, Universidade de São Paulo,
São Paulo, SP, Brasil

Corresponding author: M.S. Lauretto
E-mail: lauretto@ime.usp.br

Genet. Mol. Res. 7 (3): 883-897 (2008)
Received june 2, 2008
Accepted August 11, 2008
Published September 23, 2008

ABSTRACT. Gene clustering is a useful exploratory technique to group together genes with similar expression levels under distinct cell cycle phases or distinct conditions. It helps the biologist to identify potentially meaningful relationships between genes. In this study, we propose a clustering method based on multivariate normal mixture models, where the number of clusters is predicted via sequential hypothesis tests: at each step, the method considers a mixture model of m components ($m = 2$ in the first step) and tests if in fact it should be $m - 1$. If the hypothesis is rejected, m is increased and a new test is carried out. The method continues (increasing m) until the hypothesis is accepted. The theoretical core of the method is the full Bayesian significance test, an intuitive Bayesian approach, which needs no model complexity penalization nor positive probabilities for sharp hypotheses. Numerical experiments were based on a cDNA microarray dataset consisting of expression levels of 205 genes belonging to four functional categories, for 10 distinct strains of *Saccharomyces cerevisiae*. To analyze the method's sensitivity to data dimension, we performed principal components analysis on the original dataset and predicted the number of classes using 2 to 10

principal components. Compared to Mclust (model-based clustering), our method shows more consistent results.

Key words: Gene clustering; Mixture models; Significance test; Expression data analysis

INTRODUCTION

Clustering is a useful exploratory technique for gene expression data, as it groups similar objects together and allows the biologist to identify potentially meaningful relationships between the objects - e.g., genes, tissues, etc. (Yeung et al., 2003). In gene clustering, the interest is to group genes with similar expression levels under distinct phases of the cell cycle, or under distinct conditions (e.g., normal *vs* cancer tissues).

Several clustering methods are available in the literature and have been widely applied in expression data analysis - see below. A particular family of parametric methods based on normal mixture models shows good comparative results (Banfield and Raftery, 1993; Biernacki and Govaert, 1998; Yeung et al., 2001; Thalamuthu et al., 2006).

In this study, we propose a clustering method based on normal mixture models, where the number of clusters is decided by sequential significance tests. The starting point is to consider a mixture model with two clusters and to test if one cluster's weight is zero. If, in this first step, the null hypothesis is rejected, a new cluster must be considered. A new test of two against three clusters is performed. Again, if the null hypothesis is rejected, a fourth cluster must be considered. The theoretical core of the method is the full Bayesian significance test (FBST), an intuitive Bayesian approach that needs no model complexity penalization nor positive probabilities for sharp hypotheses. In this aspect, FBST is an intrinsic regularization criterion.

In Material and Methods section, we introduce the FBST and the theory and implementation of normal mixture models in the FBST context, and describe the method for deciding the number of mixture components. In Data Analysis section, we describe the analysis performed on a gene expression dataset, and in Discussion and Final Remarks section, we discuss the results obtained.

In the remainder of this section, we briefly introduce some classic methods widely applied to gene expression analysis - hierarchical clustering (HC), K-means, fuzzy C-means, support vector clustering, and mixture model clustering. In particular, we briefly discuss two approaches for determination of the number of clusters in mixture models.

Hierarchical clustering

This is one of the most widely used algorithms, due to its conceptual and practical simplicity. Agglomerative HC (Johnson, 1967) starts by considering the n data points as n clusters, and at each iterative stage, a pair of clusters with the shortest distance between them is joined to form a new cluster. Divisive HC starts by considering a unique cluster and at each iterative step a cluster is divided in two. In both methods, a hierarchical tree is constructed after n - 1 steps. Initially, a distance (or dissimilarity) matrix of all pairs of points must be computed, using some distance function (Euclidean, Mahalanobis, Manhattan, etc.). To define distances between two clusters, different linkages including single linkage (shortest pair-wise distance), complete linkage (largest distance) or average linkage (average distance) may be chosen.

Two difficulties with this method are its lack of robustness due to local optimization and the sensitivity of the results to different distance and linkage criteria used.

K-means

Proposed by MacQueen (1967), K-means aims to split the data into K clusters by minimizing the within cluster dispersion

$$\sum_{j=1}^{K} \sum_{x_i \in C_j} \|x_i - v_j\|^2 , \qquad \text{(Equation 1)}$$

where $C_1 \ldots C_K$ denotes the clusters, v_j is the center of cluster j and ||.|| denotes Euclidean distance. The method starts with K arbitrary cluster centers. Each step consists in labeling data points with their nearest cluster centers, and to update the centers of the new clusters. The procedure stops when the clusters formed at two consecutive steps are the same. One way to avoid the dependence on the initial centers is to run K-means algorithm multiple times with random initial centers and select the center solution with the smallest within cluster sum of squares.

Fuzzy C-means

This algorithm aims to find a partition of C clusters, which minimizes the weighted sum of within cluster distances (Dunn, 1973; Bezdek, 1981). Formally, the method minimizes the objective function

$$J = \sum_{i=1}^{n} \sum_{j=1}^{C} u_{ij}^{m} d(x_i, v_j) , \qquad \text{(Equation 2)}$$

where v_j is the center of cluster j, u_{ij} is the probability of point i belonging to cluster j, d is a function distance (usually the Euclidean distance) and $m > 1$ is the Fuzzy parameter.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster center c_j by:

$$u_{ij} = \left[\sum_{k=1}^{C} \left(\frac{d(x_i, v_j)}{d(x_i, v_k)} \right)^{2/(m-1)} \right]^{-1} , \; v_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}} . \qquad \text{(Equation 3)}$$

Support vector clustering

In support vector clustering (Ben-Hur et al., 2001), data points are mapped from data space to a high dimensional feature space using a Gaussian kernel. In feature space, the algorithm looks for the smallest sphere that encloses the image of the data. This sphere is mapped back to data space, where it forms a set of contours, which enclose the data points. These contours are interpreted as cluster boundaries, and points enclosed by each separate contour

are associated with the same cluster. As the width parameter of the Gaussian kernel is decreased, the number of disconnected in data space increases, leading to an increasing number of clusters. Since the contours can be interpreted as delineating the support of the underlying probability distribution, the algorithm can be viewed as one identifying valleys in this probability distribution.

Mixture model clustering

In a mixture model, the probability density function (pdf) is a linear combination of several pdf's; that is,

$$f(.) = w_1 f(.|\psi_1) + w_2 f(.|\psi_2) + \ldots + w_m f(.|\psi_m) \,, \qquad \text{(Equation 4)}$$

where $f(.|\psi)$ is a given parametric family of densities indexed by a parameter ψ (usually normal). In the classification context, given a dataset X, each observation is assumed to have arisen from one of m (possibly unknown) groups. The mixture proportion w_k represents the relative frequency of group k in the population. This model provides a framework from which observations may be clustered together into groups for discrimination or classification (Stephens, 1997).

One important aspect of mixture models is that the mixture parameters (cluster proportions, means and covariance structures) provide, in a compact form, the structural information of the sample space. Moreover, the classification probabilities assigned by the mixture model for the sample data are also more suitable and informative than the cluster labels provided by some traditional clustering methods.

To decide the number of components, information criteria are commonly used. Akaike (1974) information criterion (AIC) and Schwartz' Bayesian information criterion (BIC) (Schwarz, 1978; Fraley and Raftery, 1999) are the most often used, and are computed by

$$AIC = -2\lambda + 2\kappa; \; BIC = -2\lambda + \kappa log(n), \qquad \text{(Equation 5)}$$

where λ is the maximum model log-likelihood, κ its number of parameters, and n the sample size. These are regularization criteria, weighting the model fit against the number of parameters. Larger AIC and BIC scores indicate stronger evidence for the corresponding model.

One alternative approach is the reversible jump Markov Chain Monte Carlo (MCMC) method (Richardson and Green, 1997), which is capable of "jumping" between the parameter subspaces corresponding to different numbers of components in the mixture. A sample from the full joint distribution of all unknown variables is thereby generated and used as a basis for estimating the number of components.

In the family of mixture model approaches, the software Mclust (Banfield and Raftery, 1993; Fraley and Raftery, 1999) is available as an R package and has been used in many applications. In Mclust, each covariance matrix is parameterized by eigenvalue decomposition in the form

$$V^k = \lambda^k Q^k D^k (Q^k)' \,, \qquad \text{(Equation 6)}$$

where Q^k is the orthogonal matrix of eigenvectors, D^k is a diagonal matrix whose elements are proportional to the eigenvalues of V^k, and λ^k is a scalar. Q^k determines the orientation of the

principal components of V^k, D^k determines its shape and λ^k determines its volume. Different constraints over some of these parameters provide several alternative models (Celeux et al., 1996), some of them implemented in Mclust. Given a specific covariance model and number of components, the weights w_k, the mean vectors b^k and constrained covariance matrixes V^k are estimated via the expectation-maximization (EM) algorithm (described later in this study). The variance structure and the number of components are selected via BIC. In this paper, the performance of our proposed method is compared to Mclust.

MATERIAL AND METHODS

The full Bayesian significance test

The FBST was proposed by Pereira and Stern (1999) as a coherent and intuitive test. It considers the parameter space, Θ, as a subset of R^n, and the hypothesis is defined as a further restricted subset defined by vector valued inequality and equality constraints:

$$H{:}\theta\in\Theta_H \text{ where } \Theta_H=\{\theta\in\Theta\,|\,g(\theta)\leq 0\wedge h(\theta)=0\}. \qquad \text{(Equation 7)}$$

For simplicity, we often use H for Θ_H. FBST is particularly focused on precise hypotheses, with $\dim(\Theta_H)<\dim(\Theta)$. The posterior probability density function is denoted by $f(\theta)$.

The computation of the evidence measure used on the FBST is performed in two steps:

The optimization step consists of finding f^*, the maximum (supremum) of the posterior under the null hypothesis.

The integration step consists of integrating the posterior density over the tangential set T, where the posterior is higher than anywhere in the hypothesis, i.e.,

$$\mathrm{ev}^*(H) = \Pr(\theta\in T^*\,|\,x) = \int_{T^*} f(\theta)d\theta \qquad \text{(Equation 8)}$$

where:

$$T^* = \{\theta\in\Theta; f(\theta)>f^*\} \quad \text{and} \quad f^* = \sup_H f(\theta) \qquad \text{(Equation 9)}$$

The measure $\mathrm{ev}^*(H)$ is the evidence against H, and $\mathrm{ev}(H) = 1 - \mathrm{ev}^*(H)$ is the evidence supporting (or in favor of) H. Intuitively, if $\mathrm{ev}^*(H)$ is "large", T is "heavy", and the hypothesis set is in a region of "low" posterior density, meaning a "strong" evidence against H.

For a hypothesis test procedure, a critical level τ must be chosen such that H is rejected if and only if $\mathrm{ev}^*(H) > \tau$. We discuss briefly some criteria in the last subsection.

Multivariate normal mixtures

In a d-dimensional multivariate finite mixture model with m components (or classes), and sample size n, any given sample x^j is of class k with probability w_k; the weights, w_k, give the probability that a new observation is of class k. A sample j of class $k = c(j)$ is distributed with density $f(x^j \mid \psi_k)$.

This paragraph defines some general matrix notation. Let $r : s : t$ indicate either the vector $[r,r + s,r + 2s,\ldots t]$ or the corresponding index range from r to t with step s; $r: t$ is short hand for r: 1: t. A matrix array has a superscript index, like $S^1 \ldots S^m$. So $S^k_{h,i}$ is the h-row, i-column element of matrix S^k. We may write a rectangular matrix, X, with the row (or shorter range) index subscript, and the column (or longer range) index superscript. Thus, x_i, x^j, and x^j_i are row i, column j, and element (i,j) of matrix X. **0** and **1** are matrices of zeros and ones, where dimensions are given by the context. $V > 0$ is a positive definite matrix. In this paper, let h, i be indices in the range 1: d, k in 1: m, and j in 1: n.

The classifications z^j_k are boolean variables indicating whether or not x^j is of class k, i.e., $z^j_k = 1$ if $c(j) = k$. Z is not observed, being therefore called named latent variable or missing data. Conditioning on the missing data, we get:

$$\begin{aligned} f(x^j|\theta) &= \sum_{k=1}^{m} f(x^j|\theta,z^j_k) f(z^j_k|\theta) = \sum_{k=1}^{m} w_k f(x^j|\psi_k) \\ f(X|\theta) &= \prod_{j=1}^{n} f(x^j|\theta) = \prod_{j=1}^{n} \sum_{k=1}^{m} w_k f(x^j|\psi_k) \end{aligned} \qquad \text{(Equation 10)}$$

Given the mixture parameters, $\theta = \{w_1 \ldots w_m, \psi_1 \ldots \psi_m\}$ and the observed data, X, the conditional classification probabilities, $P = f(Z|X, \theta)$, are:

$$p^j_k = f(z^j_k|x^j,\theta) = \frac{f(z^j_k,x^j|\theta)}{f(x^j|\theta)} = \frac{w_k f(x^j|\psi_k)}{\sum_{k=1}^{m} w_k f(x^j|\psi_k)} \qquad \text{(Equation 11)}$$

The likelihood for the "completed" data, X,Z, is:

$$\begin{aligned} f(X,Z|\theta) &= \prod_{j=1}^{n} f(x^j|\psi_{c(j)}) f(z^j_k|\theta) \\ &= \prod_{k=1}^{m} \left[w_k^{y_k} \prod_{j|c(j)=k} f(x^j|\psi_k) \right], \end{aligned} \qquad \text{(Equation 12)}$$

where y_k is the number of samples of class k, i.e., $y_k = \Sigma_j z^j_k$ or $y = Z\,\mathbf{1}$.

We will see in the following sections that considering the missing data Z, and the conditional classification probabilities P, is the key for successfully solving the numerical integration and optimization steps of the FBST.

In this article, we will focus on Gaussian finite mixture models, where each component follows a normal density with mean b^k and variance matrix V^k, or precision $R^k = (V^k)^{-1}$:

$$f(x^j|\psi_k) = N(x^j|b^k,R^k) \qquad \text{(Equation 13)}$$

Next, we focus the theory of general mixture models on the Dirichlet-normal-Wishart case.

The Dirichlet-normal-Wishart prior

Consider the random matrix X^j_i, i in 1: d, j in 1: n, $n > d$, where each column contains a sample element from a d-multivariate normal distribution with parameters b (mean) and V

(covariance), or $R = V^{-1}$ (precision). Let u and S denote the statistics:

$$u = (1/n)\ \sum_{j=1}^{n} x^j = (1/n)X\mathbf{1}\ ,\ \ S = \sum_{j=1}^{n}(x^j - b)\otimes(x^j - b)' = (X-b)(X-b)' \quad \text{(Equation 14)}$$

The random vector u has a normal distribution with mean b and precision nR. The random matrix S has a Wishart distribution with n degrees of freedom and precision matrix R. The normal, Wishart and normal-Wishart pdfs have the expressions:

$$\begin{aligned} N(u\,|\,n,b,R) &= \left(\frac{n}{2\pi}\right)^{d/2}|R|^{1/2}\exp\left(-(n/2)(u-b)'R(u-b)\right) \\ W(S\,|\,e,R) &= c^{-1}\,|S|^{(e-d-1)/2}\exp\left(-(1/2)\mathrm{tr}(SR)\right) \end{aligned} \quad \text{(Equation 15)}$$

with normalization constant

$$c = |R|^{-e/2}\,2^{ed/2}\,\pi^{d(d-1)/4}\ \prod_{i=1}^{d}\Gamma((e-i+1)/2). \quad \text{(Equation 16)}$$

Now consider the matrix X as above, with unknown mean b and unknown precision matrix R, and the statistic

$$S = \sum_{j=1}^{n}(x^j - u)\otimes(x^j - u)' = (X-u)(X-u)'. \quad \text{(Equation 17)}$$

The conjugate family of priors for multivariate normal distributions is the normal-Wishart (DeGroot, 1970). For the precision matrix R, take as prior the Wishart distribution with $\dot{e} > d - 1$ degrees of freedom and precision matrix $\dot{S}$ and, given R, take as prior for b a multivariate normal with mean $\dot{u}$ and precision $\dot{n}R$, i.e., let us take the normal-Wishart prior $NW(b, R\,|\,\dot{n},\dot{e},\dot{u},\dot{S})$. Then, the posterior distribution for R is a Wishart distribution with $\ddot{e}$ degrees of freedom and precision $\ddot{S}$, and the posterior for b, given R, is k-normal with mean $\ddot{u}$ and precision $\ddot{n}R$, i.e., we have the normal-Wishart posterior:

$$NW(b,R\,|\,\ddot{n},\ddot{e},\ddot{u},\ddot{S}) = W(R\,|\,\ddot{e},\ddot{S})\ N(b\,|\,\ddot{n},\ddot{u},R)$$

$$\ddot{n}=\dot{n}+n\ ,\ \ \ddot{e}=\dot{e}+n\ ,\ \ \ddot{u}=(nu+\dot{n}\dot{u})/\ddot{n} \quad \text{(Equation 18)}$$

$$\ddot{S}=S+\dot{S}+(n\dot{n}/\ddot{n})(u-\dot{u})\otimes(u-\dot{u})'.$$

All covariance and precision matrices are supposed to be positive definite, and proper priors have $\dot{e} \geq d$ and $\dot{n} \geq 1$. Non-informative normal-Wishart improper priors are given by $\dot{n} = 0$, $\dot{u} = 0$, $\dot{e} = 0$, $\dot{S} = 0$, i.e., we take a Wishart with 0 degrees of freedom as prior for R, and a constant prior for b (DeGroot, 1970). Then, the posterior for R is a Wishart with n degrees of freedom and precision S, and the posterior for b, given R, is d-normal with mean u and precision nR.

The conjugate prior for a multinomial distribution is a Dirichlet distribution:

$$M(y\,|\,n,w) = \left(n!/y_1!\ldots y_m!\right) w_1^{\,y_1}\ldots w_m^{\,y_m}$$
$$D(w\,|\,y) = \left(\Gamma(y_1+\ldots+y_k)/\Gamma(y_1)\ldots\Gamma(y_k)\right) \prod_{k=1}^{m} w_k^{\,y_k-1} \qquad \text{(Equation 19)}$$

with $w > \mathbf{0}$ and $w\mathbf{1} = 1$. Prior information given by $\dot{y}$, and observation y, result in the posterior parameter $\ddot{y} = \dot{y} + y$. A non-informative prior is given by $\dot{y} = \mathbf{1}$.

Finally, for the Gaussian mixture model with Dirichlet-normal-Wishart prior, we can write the posterior

$$f(\theta\,|\,X,\dot{\theta}) \propto f(X\,|\,\theta) f(\theta\,|\,\dot{\theta})$$
$$f(X\,|\,\theta) = \prod_{j=1}^{n} \sum_{k=1}^{m} p_k^j w_k N(x^j\,|\,b^k,R^k)$$
$$f(\theta\,|\,\dot{\theta}) = D(w\,|\,\dot{y}) \prod_{k=1}^{m} NW(b^k,R^k\,|\,\dot{n}_k,\dot{e}_k,\dot{u}^k,\dot{S}^k) \qquad \text{(Equation 20)}$$
$$p_k^j = \frac{w_k N(x^j\,|\,b^k,R^k)}{\sum_{k=1}^{m} w_k N(x^j\,|\,b^k,R^k)}$$

and completed posterior:

$$f(\theta|X,Z,\dot{\theta}) \propto f(\theta|X,Z) f(\theta|\dot{\theta}) = D(w|\ddot{y}) \prod_{k=1}^{m} NW(b^k,R^k\,|\,\ddot{n}_k,\ddot{e}_k,\ddot{u}^k,\ddot{S}^k)$$
$$y = Z\mathbf{1}\ ,\ \ddot{y} = \dot{y}+y\ \ ,\ \ddot{n} = \dot{n}+y\ ,\ \ddot{e} = \dot{e}+y,\ ,$$
$$u^k = (1/y_k)\sum_{j=1}^{n} z_k^j x^j\ ,\ S^k = \sum_{j=1}^{n} z_k^j (x^j-u^k)\otimes(x^j-u^k)'\ , \qquad \text{(Equation 21)}$$
$$\ddot{u}^k = (1/\ddot{y}_k)(\dot{n}_k\dot{u}^k+y_k u^k)\ ,\ \ddot{S}^k = S^k+\dot{S}^k+(\dot{n}_k y_k/\ddot{n}_k)(u^k-\dot{u}^k)\otimes(u^k-\dot{u}^k)'$$

Gibbs sampling

In order to integrate a function over the posterior measure, we use an ergodic Markov chain. The form of the chain below is known as Gibbs sampling, and its use for numerical integration is known as Markov Chain Monte Carlo or MCMC.

At each iteration, given θ, we can compute P. Given P, we draw z^j by means of multinomial distribution $f(z^j\,|\,p^j)$. Given the latent variables, Z, we have simple conditional posterior density expressions for the mixture parameters:

$$f(w\,|\,Z,\dot{y}) = D(w\,|\,\ddot{y})$$

$$f(R^k\,|\,X,Z,\dot{e}_k,\dot{S}^k) = W(R\,|\,\ddot{e}_k,S^k) \qquad \text{(Equation 22)}$$

$$f(b^k\,|\,X,Z,R^k,\dot{n}_k,\dot{u}^k) = N(b\,|\,\ddot{n}_k,\ddot{u}^k,R^k)$$

Gibbs sampling is the MCMC generated by cyclically updating variables Z, θ, and P, by drawing θ and Z from the above distributions (Gilks et al., 1996; Häggström, 2002). A multinomial variate can be drawn using a uniform generator. A Dirichlet variate w can be drawn using a gamma generator with shape and scale parameters α and β (Gentle, 1998). Johnson (1987) describes a simple procedure to generate the Cholesky factor of a Wishart variate $W = U' U$ with n degrees of freedom, from the Cholesky factorization of the covariance parameter $V = R^{-1} = C' C$, and a chi-square generator:

$$\begin{gathered} g_k = G(y_k, 1); \\ w_k = \frac{g_k}{\sum_{k=1}^{m} g_k}; \\ \text{For } i < j,\ B_{ij} = \mathrm{N}(0,1); \\ B_{i,i} = \sqrt{\chi^2(n-i+1)}; \\ U = BC. \end{gathered} \qquad \text{(Equation 23)}$$

For good performance, all subsequent matrix computations proceed directly from Cholesky factors (Golub and van Loan, 1989; Jones, 1985).

Given a mixture model, we obtain an equivalent model renumbering the components 1: m by a permutation σ ([1: m]). This symmetry must be broken in order to have an identifiable model (Stephens, 1997). Let us assume that there is an order criterion that can be used when numbering the components. If the components are not in the correct order, label switching is the operation of finding permutation σ ([1: m]) and renumbering the components, so that the order criterion is satisfied.

If we want to look consistently at the classifications produced during an MCMC run, we must enforce a label switching to break all non-identifiability symmetries. For example, in the Dirichlet-normal mixture model, we could choose to order the components (switch labels) according to the rank given by:

1. A given linear combination of the vector means, $c' b^k$;
2. The variance determinant $|V^k|$.

The choice of a good label switching criterion should consider not only the model structure and the data, but also the semantics and interpretation of the model.

The semantics and interpretation of the model may also dictate that some states, like certain configurations of the latent variables Z, are either meaningless or invalid, and shall not be considered as possible solutions. The MCMC can be adapted to deal with forbidden states by implementing rejection rules that prevent the chain from entering the forbidden regions of the complete and/or incomplete state space (Bennett, 1976; Meng and Wong, 1996).

EM algorithm for maximum likelihood and maximum *a posteriori* estimation

The EM algorithm optimizes the log-posterior function $fl(X \mid \theta) + fl(\theta \mid \dot{\theta})$ (Dempster et al., 1977; Ormoneit and Tresp, 1995). The EM is derived from the conditional log-likelihood, and the Jensen inequality:

$$\text{If } w,y > \mathbf{0},\ w'\mathbf{1} = 1 \text{ then } \log w'y \geq w'\log y. \qquad \text{(Equation 24)}$$

Let θ and $\tilde{q}$ be our current and next estimate of the maximum *a posteriori*, and $p^j_k = f(z^j_k \mid x^j, \theta)$ the conditional classification probabilities. At each iteration, the log-posterior improvement is:

$$\delta(\tilde{\theta},\theta \mid X,\dot{\theta}) = fl(\tilde{\theta} \mid X,\dot{\theta}) - fl(\theta \mid X,\dot{\theta}) = \delta(\tilde{\theta},\theta \mid X) + \delta(\tilde{\theta},\theta \mid \dot{\theta})$$

$$\delta(\tilde{\theta},\theta \mid \dot{\theta}) = fl(\tilde{\theta} \mid \dot{\theta}) - fl(\theta \mid \dot{\theta})$$

$$d(\tilde{q},q \mid X) = fl(X \mid \tilde{q}) - fl(X \mid q) = \sum_j d(\tilde{q},q \mid x^j) \qquad \text{(Equation 25)}$$

$$\delta(\tilde{\theta},\theta \mid x^j) = fl(x^j \mid \tilde{\theta}) - fl(x^j \mid \theta) = \log \sum_k \tilde{w}_k f(x^j \mid \tilde{\psi}_k) - fl(x^j \mid \theta) =$$

$$= \log \sum_k \frac{p^j_k}{p^j_k} \frac{\tilde{w}_k f(x^j \mid \tilde{\psi}_k)}{f(x^j \mid \theta)} \geq \Delta(\tilde{\theta},\theta \mid x^j) = \sum_k p^j_k \log \frac{\tilde{w}_k f(x^j \mid \tilde{\psi}_k)}{p^j_k f(x^j \mid \theta)} .$$

Hence,

$$\Delta(\tilde{\theta},\theta \mid X,\dot{\theta}) = \Delta(\tilde{\theta},\theta \mid X) + \delta(\tilde{\theta},\theta \mid \dot{\theta}) \qquad \text{(Equation 26)}$$

is a lower limit to $\delta(\tilde{\theta},\theta \mid X,\dot{\theta})$. Also,

$$\Delta(\theta,\theta \mid X,\dot{\theta}) = \delta(\theta,\theta \mid X,\dot{\theta}) = 0. \qquad \text{(Equation 27)}$$

Thus, under mild differentiability conditions, both surfaces are tangent, assuring convergence of EM to the nearest local maximum. However, maximizing $\Delta(\tilde{\theta},\theta \mid X,\dot{\theta})$ over $\tilde{\theta}$ is the same as maximizing

$$Q(\tilde{\theta},\theta) = \sum_{k,j} p^j_k \log\left(\tilde{w}_k f(x^j \mid \tilde{\psi}_k)\right) + fl(\tilde{\theta} \mid \dot{\theta}) \qquad \text{(Equation 28)}$$

and each iteration of the EM algorithm breaks down in two steps:

E-step: compute $P = E(Z \mid X,\theta)$.
M-step: optimize $Q(\tilde{\theta},\theta)$, given P.

For the Gaussian mixture model, with a Dirichlet-normal-Wishart prior,

$$Q(\tilde{\theta},\theta) = \sum_{k=1}^{m} \sum_{j=1}^{n} p^j_k \left(\log \tilde{w}_k + \log N(x^j \mid \tilde{b}^k, \tilde{R}^k)\right) + fl(\tilde{\theta} \mid \dot{\theta})$$
$$fl(\tilde{\theta} \mid \dot{\theta}) = \log D(\tilde{w} \mid \dot{y}) + \sum_{k=1}^{m} \log NW(\tilde{b}^k, \tilde{R}^k \mid \dot{n}_k, \dot{e}_k, \dot{u}^k, \dot{S}^k) \qquad \text{(Equation 29)}$$

For normal mixture models, the E-step is performed as explained in Equation 20. The Lagrange optimality conditions give an analytical solution for the M-step:

$$y = P\mathbf{1}\ ,\ \tilde{w}_k = \frac{y_k + \dot{y}_k - 1}{n - m + \sum_{k=1}^{m} \dot{y}_k}$$

$$u^k = \frac{1}{y_k} \sum_{j=1}^{n} p_k^j x^j\ ,\ S^k = \sum_{j=1}^{n} p_k^j (x^j - \tilde{b}^k) \otimes (x^j - \tilde{b}^k)' \qquad \text{(Equation 30)}$$

$$\tilde{b}^k = \frac{\dot{n}_k \dot{u}^k + y_k u^k}{\dot{n}_k + y_k}\ ,\ \tilde{V}^k = \frac{S^k + \dot{n}_k (\tilde{b}^k - \dot{u}^k) \otimes (\tilde{b}^k - \dot{u}^k)' + \dot{S}^k}{y_k + \dot{e}_k - d}$$

In more general (non-Gaussian) mixture models, if an analytical solution for the M-step is not available, a robust local optimization algorithm can be used, for example, Martínez (2000).

The EM is a local optimizer, but the MCMC provides many starting points, so that we have the basic elements for a global optimizer. To avoid using many starting points going to a same local maximum, we can filter the top portion of the MCMC output (i.e., points of highest posterior densities) using a clustering algorithm, and select a starting point from each cluster. For better efficiency, or more complex problems, the stochastic EM algorithm can be used to provide starting points near each important local maximum (Celeux et al., 1996; Pflug, 1996; Spall, 2003). Numerical integration and global optimization are relatively costly. The average computing time of a single FBST e-value on the mixture models discussed in this paper is about 1 min on a Pentium 4 personal computer. In contrast, Mclust (see Introduction) takes about 3 s. Nevertheless, one should note that Mclust is a well-developed and optimized production software, while we are using a naively programmed prototype. Although slower than most heuristic algorithms, including Mclust, we do not see this computing time as an impediment to data analysis. Nevertheless, heuristic algorithms may be more appropriate for the exploration of large databanks, while our procedure could be used for a more refined final analysis.

Choice of number of clusters

A procedural description of the method used to decide the number of components in the mixture is given next.

Step 0: Set $m \leftarrow 2$
Step 1: Consider the mixture model

$$f(x \mid \theta) = \sum_{k=1}^{m} w_k f(x \mid \psi_k) \qquad \text{(Equation 31)}$$

and test the hypothesis $H: w_m = 0$.

Step 2: If H is not rejected, $ev^*(H) \leq \tau$, stop testing and consider the model as having m - 1 components.

If H is rejected, $ev^*(H) > \tau$, it means that the model must have at least m components. In this case, increment m by 1 and go to Step 1.

As with any significance test, this procedure requires the choice of a significance threshold level, τ. Several alternative methods have been developed for establishing this threshold:

• An asymptotically consistent threshold for a given confidence level was given by Stern (2005) and Borges and Stern (2007).

• An empirical power analysis, developed by Stern and Zacks (2002) and Lauretto et al. (2003, 2005), provides critical levels that are consistent and also effective for small samples.

• A threshold based on reference sensitivity analysis and paraconsistent logic is given by Stern (2004).

• Varuzza et al. (2008) relate the e-value threshold to standard p-value thresholds.

• Finally, Madruga et al. (2001) prove the existence of a loss function that renders the FBST a true Bayesian decision-theoretic procedure. We use this framework to justify setting $\tau = 0.5$.

DATA ANALYSIS

The present case study was based on a cDNA microarray dataset planned and produced by Ideker et al. (1985). The authors explored the process of galactose (GAL) utilization in the yeast *Saccharomyces cerevisiae*. They applied 10 initial perturbations to the GAL pathway. Wild-type and nine genetically altered yeast strains were examined, each with a complete deletion of a gene involved in GAL processing. Global changes in mRNA expression resulting from each perturbation were examined with DNA microarrays of approximately 6200 genes. In each experiment, fluorescently labeled cDNA from a perturbed strain was hybridized against labeled cDNA from the reference (wild) strain. For robustness, four replicate hybridizations were performed for each perturbation.

From these data, Yeung et al. (2003) selected a subset of 205 genes that are reproducibly measured, whose expression patterns reflect four functional categories in the Gene Ontology listings (Ashburner et al., 2000) and that are expected to be clustered together.

In our study, the input data is a matrix Y of dimension 205×10, where $Y_{i,j}$ corresponds to the average expression level of gene i under condition j.

Figure 1 shows the expression profiles of the selected genes in the 10 strains. The colors distinguish the four functional classes of genes.

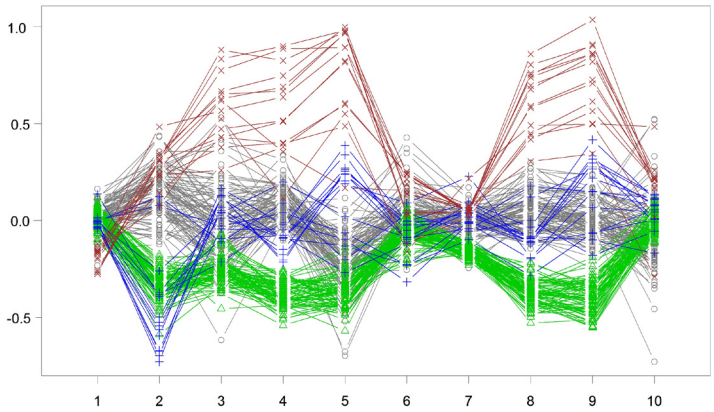

Figure 1. Expression profiles of 205 genes selected in 10 distinct strains, identified in four functional classes.

Due to the high level of noise and redundancy (distinct strains with very similar expression profiles), we applied principal components analysis as a filter for dimensionality reduction (Everitt and Dunn 1992). The cumulative variance in the principal components is presented in Table 1.

Table 1. Yeast cell data: cumulative variance on principal components (PC).

PC	1	2	3	4	5	6	7	8	9	10
% Variation	75.5	88.8	92.5	95.0	96.8	98.1	98.8	99.4	99.7	100

We performed a comparative analysis of FBST and Mclust (see Introduction), where we applied both methods on the original dataset and on the transformed data with 2 to 10 principal components, using the four functional categories as our external knowledge. The performance of algorithms was evaluated by two criteria:

• Correct prediction of the number of classes;

• Percentage of genes classified correctly according to their functional classes.

Table 2 presents the numerical results for this analysis, where each row contains the number of principal components used in data, the predicted number of clusters and the percentage of genes classified correctly for each method. The last row contains the number of predicted classes and classification precision for the original dataset.

Table 2. FBST and Mclust predicted number of clusters and percentage of genes classified correctly, according to the number of principal components (last row refers to original dataset).

Principal components	FBST		Mclust	
	Predicted clusters	% Correct classification	Predicted clusters	% Correct classification
2	5	97	4	96
3	4	99	4	97
4	4	97	5	84
5	4	84	7	60
6	4	83	7	59
7	3	91	6	68
8	3	90	6	63
9	3	87	7	64
10	3	88	7	61
Original	3	89	3	92

FBST = full Bayesian significance test; Mclust = model-based clustering.

DISCUSSION AND FINAL REMARKS

In the first numerical experiments (results in Table 2), the clustering method based on FBST predicts the correct number of clusters in four situations (with 3, 4, 5, and 6 principal components), while Mclust provides the correct answer only in two situations (with 2 and 3 principal components). Moreover, the clustering based on FBST shows a stable behavior. That is, in the presence of little information (2 principal components), FBST predicts more clusters than the expected (5). As more relevant information is added to the data (3 and 4 principal components), FBST provides the correct number of clusters and a high percentage of genes

classified correctly. On the other hand, the predicted number of clusters in Mclust oscillates as the number of principal components increases. That is, the behavior of Mclust is very sensitive to redundancy and noise.

Also notice that the data filtered with 10 principal components are just the original data in transformed coordinates. Since the FBST is an invariant procedure, the numerical results are the same, up to numerical precision, for the original and transformed coordinate systems. In contrast, Mclust, which is a non-invariant procedure, is highly sensitive to this change of coordinates, yielding quite different results.

These results suggest that our approach is robust for model selection and deserves further development and investigation for gene expression clustering. Some possible refinements are the analysis of alternative priors, the implementation of stochastic EM or the integration of the MCMC and the optimization procedures for a better computational performance.

ACKNOWLEDGMENTS

The authors are grateful to CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo, and the University of São Paulo Program on Bioinformatics for financial support.

REFERENCES

Akaike H (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* 19: 716-723.

Ashburner M, Ball CA, Blake JA, Botstein D, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25: 25-29.

Banfield JD and Raftery AE (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49: 803-821.

Ben-Hur A, Horn D, Sielgmann HT and Vapnik V (2001). Support vector clustering. *J. Mach. Learn. Res.* 2: 125-137.

Bennett CH (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* 22: 245-268.

Bezdek JC (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Biernacki C and Govaert G (1998). Choosing Models in Model-based Clustering and Discriminant Analysis. Technical Report 3509, INRIA, Montbonnot-Saint-Martin.

Borges W and Stern JM (2007). The rules of logic composition for the Bayesian epistemic e-values. *Logic J. IGPL* 15: 401-420.

Celeux G, Chauveau D and Diebolt J (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *J. Stat. Comput. Simul.* 55: 287-314.

DeGroot MH (1970). Optimal Statistical Decisions. Wiley, New York.

Dempster AP, Laird NM and Rubim DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* 39: 1-38.

Dunn JC (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.* 3: 32-57.

Everitt BS and Dunn G (1992). Applied Multivariate Data Analysis. Oxford University Press, New York.

Fraley C and Raftery AE (1999). MCLUST: Software for model-based cluster analysis. *J. Classif.* 16: 297-306.

Gentle JE (1998). Random Number Generation and Monte Carlo Methods. Springer, New York.

Gilks WR, Richardson S and Spiegelhalter DJ (1996). Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics. CRC Press, New York.

Golub GH and van Loan CF (1989). Matrix Computations. Johns Hopkins University Press, Baltimore.

Häggström O (2002). Finite Markov Chains and Algorithmic Applications. Cambridge University Press, Cambridge.

Ideker T, Thorsson V, Ranish JA, Christmas R, et al. (1985). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292: 511-514.

Johnson ME (1987). Multivariate Statistical Simulation. Wiley, New York.

Johnson SC (1967). Hierarchical clustering schemes. *Psychometrika* 32: 241-254.

Jones MC (1985). Generating inverse wishart matrices. *Comm. Stat. Simul. Comput.* 14: 511-514.
Lauretto MS and Stern JM (2005). Testing Significance in Bayesian Classifiers. Vol. 132. In: Frontiers in Artificial Intelligence and Applications (Nakamatsu K and Abe JM, eds.). IOS Press, Amsterdam.
Lauretto MS, Pereira CAB, Stern JM and Zacks S (2003). Full Bayesian significance test applied to multivariate normal structure models. *Braz. J. Probab. Stat.* 17: 147-168.
Madruga RM, Esteves LG and Wechsler S (2001). On the Bayesianity of Pereira-Stern tests. *Test* 10: 291-299.
Martínez JM (2000). Box-quacan and the implementation of augmented Lagrangian algorithms for minimization with inequality constraints. *Comput. Appl. Math.* 19: 31-56.
McQueen JB (1967). Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Stat. Probab.* 1: 281-297.
Meng XL and Wong WH (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* 6: 831-860.
Ormoneit D and Tresp V (1995). Improved Gaussian mixture density estimates using Bayesian penalty terms and network averaging. *Adv. Neural Inform. Proc. Systems* 8: 542-548.
Pereira CAB and Stern JM (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy* 1: 99-110.
Pflug GC (1996). Optimization of Stochastic Models: The Interface Between Simulation and Optimization. Kluwer Academic, Boston.
Richardson S and Green PJ (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Royal Stat. Soc. B* 59: 731-758.
Schwarz G (1978). Estimating the dimension of a model source. *Ann. Stat.* 6: 461-464.
Spall JC (2003). Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley, Hoboken.
Stephens M (1997). Bayesian Methods for Mixtures of Normal Distributions. PhD thesis, University of Oxford, Oxford.
Stern JM (2004). Inconsistency Analysis for Statistical Tests of Hypotheses. In: Soft Methodology and Random Information Systems (López-Díaz M and Gil MA, eds.). Springer, New York, 267-274.
Stern JM (2005). Cognitive Constructivism, Eigen-Solutions and Sharp Statistical Hypotheses. In: Foundations of Information Science (Petitjean M, ed.). MDPI, Basel, 1-23.
Stern JM and Zacks S (2002). Testing the independence of Poisson variates under the Holgate bivariate distribution: the power of a new evidence test. *Stat. Probab. Lett.* 60: 313-320.
Thalamuthu A, Mukhopadhyay I, Zheng X and Tseng GC (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* 22: 2405-2412.
Varuzza L, Gruber A and Pereira CAB (2008). Significance tests for comparing digital gene expression profiles. In: Nature Precedings. Available at http://precedings.nature.com/documents/2002/version/3.
Yeung KY, Fraley C, Murua A, Raftery AE, et al. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17: 977-987.
Yeung KY, Medvedovic M and Bumgarner RE (2003). Clustering gene-expression data with repeated measurements. *Genome Biol.* 4: R34.