A Hierarchical Mixture Model for Gene Expression Data

Luisa Scaccia[1] and Francesco Bartolucci[2]

[1] Dipartimento di Scienze Statistiche,
Università degli Studi di Perugia, Italy
luisa@stat.unipg.it
[2] Istituto di Scienze Economiche,
Università di Urbino "Carlo Bo", Italy
Francesco.Bartolucci@uniurb.it

Abstract. We illustrate the use of a mixture of multivariate Normal distributions for clustering genes on the basis of Microarray data. We follow a hierarchical Bayesian approach and estimate the parameters of the mixture using Markov chain Monte Carlo (MCMC) techniques. The number of components (groups) is chosen on the basis of the Bayes factor, numerically evaluated using the Chib and Jelaizkov (2001) method. We also show how the proposed approach can be easily applied in recovering missing observations, which generally affect Microarray data sets. An application of the approach for clustering yeast genes according to their temporal profiles is illustrated.

1 Introduction

Microarray experiments consist in recording the expression levels of thousands of genes under a wide set of experimental conditions. The expression of a gene is defined as its transcript abundance, i.e. the frequency with which the gene is copied to induce, for example, the synthesis of a certain protein. One of the main aims of researchers is clustering genes according to similarities between their expression levels across conditions. A wide range of statistical methods (see Yeung et al. (2001) for a review) have been proposed for this purpose. Standard partitioning or hierarchical clustering algorithms have been successfully applied by a variety of authors (see, for instance, Spellman et al. (1998) and Tavazoie et al. (1999)) in order to identify interesting gene groups and characteristic expression patterns. However, the heuristic basis of these algorithms is generally considered unsatisfactory.

Microarray data are affected by several sources of error and often contain missing values. Outcomes of standard clustering algorithms can be very sensitive to anomalous observations and the way missing ones are imputed. A second generation of studies (see, for example, Brown et al. (2000) and Hastie et al. (2000)) sought further progress through more sophisticated and ad-hoc clustering strategies, employing resampling schemes, topology-constrained and/or supervised versions of partitioning algorithms, and "fuzzy" versions

of partitioning algorithms that can perform particularly well in the absence of clear-cut "natural" clusters. Recently, an increasing interest has been devoted to the model-based approach in which the data are assumed to be generated from a finite mixture (Fraley and Raftery (1998)). The main advantage is represented by straightforward criteria for choosing the number of components (groups) and imputing missing observations.

In this paper we show how Bayesian hierarchical mixture models may be effectively used to cluster genes. As in Yeung et al. (2001), we assume that the components of the mixture have multivariate Normal distribution with possibly different shape, location and dimension. An important issue is the choice of the number of components. We use the Bayes factor (Kass and Raftery (1995)), numerically computed through the Chib and Jelaizkov (2001) approach, as a selection criterion. We also outline how our approach may be used to recover missing data, which are frequent in Microarray datasets. Details on the model are given in Section 2. In Section 3, we describe the Bayesian estimation of the parameters, while in Section 4 we illustrate the model selection problem. Finally, in Section 5, we present an application of the proposed approach to the analysis of a Microarray study performed to identify groups of yeast genes involved in the cell cycle regulation.

2 The model

Let S be the number of experimental conditions and $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_S \end{pmatrix}'$ be the vector of the corresponding expression levels for a gene. We assume that the distribution of such a vector is a mixture of Normal distributions, that is

$$\mathbf{x} \sim \sum_{k=1}^{K} \pi_k N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where K is the number of components of the mixture, $\boldsymbol{\mu}_k$ is the mean of the k-th component, $\boldsymbol{\Sigma}_k$ its variance-covariance matrix and π_k its weight. In a Bayesian context, we also assume that:

- the number of components K is a priori unknown and uniformly distributed in the interval $[1;\ K_{\max}]$, where $K_{\max}$ is a suitable integer;
- the vector $\boldsymbol{\pi} = \begin{pmatrix} \pi_1 & \cdots & \pi_K \end{pmatrix}'$ has Dirichlet distribution with parameters $\beta_1, \ldots, \beta_K$;
- the $\boldsymbol{\mu}_k$'s are independent and have Normal distribution $N(\boldsymbol{\nu}, \boldsymbol{\Omega})$;
- the $\boldsymbol{\Sigma}_k$'s are independent and have inverse Wishart distribution $IW(\boldsymbol{\Xi}, \upsilon)$ where $\boldsymbol{\Xi}$ is an $S \times S$ symmetric, positive definite scale matrix, and υ is a precision parameter;
- $\boldsymbol{\nu}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Xi}$ and υ have noninformative improper prior (Jeffreys (1939)) with density $f(\boldsymbol{\nu}) = 1$, $f(\boldsymbol{\Omega}) = 1$, $f(\boldsymbol{\Xi}) = 1$ and $f(\upsilon) = 1$, $\forall\ \boldsymbol{\nu}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Xi}$ and υ.

This setting gives rise to the hierarchical model presented in Figure 1, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote, respectively, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ and $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$. We follow the

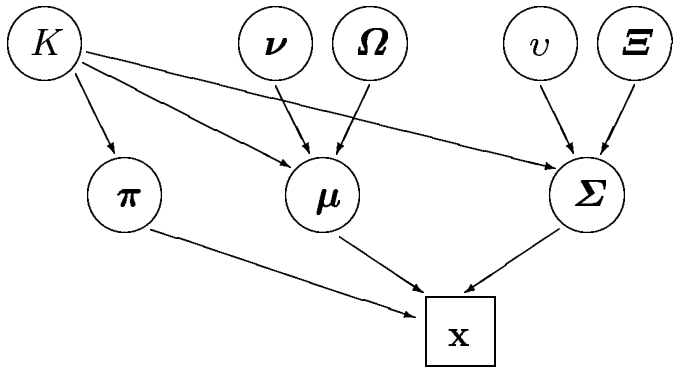

Fig. 1. Directed acyclic graph for the hierarchical mixture model.

usual convention that square boxes represent fixed or observed quantities and circles represent unknowns.

3 Bayesian estimation

3.1 Bayesian estimation without missing data

Let $\mathbf{X}$ be the $n \times S$ data matrix, where n is the number of genes. The complexity of the mixture model presented here requires MCMC methods to approximate the joint posterior distribution of the parameters. For computational reason, we introduce the latent allocation variables $\mathbf{z} = (z_1 \cdots z_n)$, where z_i indicates the component to which the i-th gene belongs; note that $p(z_i = k) = \pi_k$ a priori. Conditionally on $\mathbf{z}$, the observations $\mathbf{x}_i$'s are independent with conditional distribution $N(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$, given $\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.

For a fixed number K of components, the sampler we consider performs the following operations for a suitable number of times, T, after allowing for a burn-in period:

- update $\boldsymbol{\nu}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Xi}$ and υ, in turn, through separate Metropolis-Hastings steps. For example, to update $\boldsymbol{\nu}$ we draw $\boldsymbol{\nu}^*$ from an appropriate proposal distribution $q(\boldsymbol{\nu}^*|\boldsymbol{\nu})$ and accept it as the new value of the parameter vector with probability

$$\alpha(\boldsymbol{\nu}, \boldsymbol{\nu}^*) = \min\left\{1, \frac{p(\boldsymbol{\mu}|\boldsymbol{\nu}^*, \boldsymbol{\Omega})q(\boldsymbol{\nu}|\boldsymbol{\nu}^*)}{p(\boldsymbol{\mu}|\boldsymbol{\nu}, \boldsymbol{\Omega})q(\boldsymbol{\nu}^*|\boldsymbol{\nu})}\right\}.$$

  $\boldsymbol{\Omega}$, $\boldsymbol{\Xi}$ and υ are updated in a similar way.
- update $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\pi}$ and $\mathbf{z}$, in turn, through separate Gibbs steps. For example, to update $\boldsymbol{\mu}$, we draw, independently for each k, a new value $\boldsymbol{\mu}_k^*$ from the full conditional distribution of $\boldsymbol{\mu}_k$ given all the other parameters:

$$\boldsymbol{\mu}_k | \cdots \sim N\left(\tilde{\boldsymbol{\nu}}, \tilde{\boldsymbol{\Omega}}\right),$$

where $\tilde{\boldsymbol{\Omega}} = (\boldsymbol{\Sigma}_k^{-1} n_k + \boldsymbol{\Omega}^{-1})^{-1}$ and $\tilde{\boldsymbol{\nu}} = \tilde{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_k^{-1} \sum_{i:z_i=k} \mathbf{x}_i + \boldsymbol{\Omega}^{-1}\boldsymbol{\nu})$, with n_k being the number of genes currently allocated to the k-th group. The new parameter value $\boldsymbol{\mu}_k^*$ is accepted with probability 1. $\boldsymbol{\Sigma}$, $\boldsymbol{\pi}$ and $\mathbf{z}$ are updated in a similar way, drawing their values from the corresponding full conditional distributions.

The main purpose of inference, here, is to estimate the posterior membership probabilities $p(z_i = k|\mathbf{x}_i)$. These can be estimated from the MCMC output as

$$\hat{p}(z_i = k|\mathbf{x}_i) = \sum_{t=1}^{T} \delta(z_i^{(t)} = k)/T$$

where $z_i^{(t)}$ is the value of z_i at sweep t and $\delta(\cdot)$ denotes the indicator function. Membership probabilities provide a *soft* or *fuzzy* partition in which genes may not be univocally assigned to one component. However, it is possible to derive a standard (hard) partition by assigning each gene to the component which maximizes the membership probability. By averaging over the sweeps, we can also obtain estimates of the parameters of the model. For instance, the means of the clusters can be estimated as $\hat{\boldsymbol{\mu}}_k = \sum_{t=1}^{T} \boldsymbol{\mu}_k^{(t)}/T$.

3.2 Bayesian estimation with missing data

In missing data problems, both the parameters and the missing values are unknown. Since their joint posterior distribution is typically intractable, we can simulate from it iteratively, through the data augmentation (DA) algorithm: we sample from the distribution of the missing values, conditional on the current value of the parameters, and then we sample from the distribution of the parameters, conditional on the value imputed to the missing observations. Let us split $\mathbf{x}_i$ into two subvectors, $\mathbf{x}_i^o$ and $\mathbf{x}_i^u$, which refer, respectively, to the observed and unobserved expression levels for gene i. Let also $\mathbf{X}_o$ and $\mathbf{X}_u$ denote, respectively, the observed and unobserved expression levels for all the n genes. The DA algorithm consists in iterating the following steps:

I-step (imputation step): given the current values $\mathbf{z}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}$ of the parameters, draw a new value $\mathbf{X}_u^{(t+1)}$ for the missing observations from its conditional predictive distribution $p(\mathbf{X}_u|\mathbf{X}_o, \mathbf{z}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$. This is straightforward since, for each i, $\mathbf{x}_i^u$ can be drawn independently from a $N(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$, conditioned on $\mathbf{x}_i^o$.

P-step (posterior step): given $\mathbf{X}_u^{(t+1)}$, draw $\mathbf{z}^{(t+1)}$, $\boldsymbol{\mu}^{(t+1)}$ and $\boldsymbol{\Sigma}^{(t+1)}$ from their complete data posterior $p(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}_o, \mathbf{X}_u^{(t+1)})$ as within the sampler described in Section 3.1.

As before, estimates of the missing data, as well as of the parameters of the model, can be obtained by averaging over the sweeps of the algorithm

(Tanner and Wong (1987)), e.g.:

$$\hat{\mathbf{X}}_u = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_u^{(t)}.$$

4 Model selection

To select the number of components we make use of the *Bayes factor* (BF). Denote by M_K the mixture model at issue when K components are used and by $p(K)$ its prior probability. The BF between two models, say M_K and M_L, is defined as

$$B_{LK} = \frac{p(\mathbf{X}|L)}{p(\mathbf{X}|K)} \qquad \text{or, equivalently,} \qquad B_{LK} = \frac{p(L|\mathbf{X})}{p(K|\mathbf{X})} \Big/ \frac{p(L)}{p(K)}$$

where $p(\mathbf{X}|K)$ and $p(K|\mathbf{X})$ are, respectively, the *marginal likelihood* and posterior probability of model M_K (Kass and Raftery (1995)). The larger is B_{LK}, the greater is the evidence provided by the data in favor of M_L.

Direct computation of the BF is almost always infeasible and different algorithms have been proposed to estimate it. For example, the well-known Reversible Jump (RJ) algorithm (Green (1995)), which draws samples from the joint posterior distribution of the number of components and model parameters, allows to estimate $p(K|\mathbf{X})$ as the proportion of times the algorithm visited model M_K. However, when dealing with so many observations as in a typical Microarray study, RJ is expected to perform badly as the posterior distribution of the parameters is likely to be very peaked and this makes it hard to jump from one model to another. Therefore, we follow the approach of Chib and Jelaizkov (2001). They show that the marginal likelihood of each model can be obtained as the product of the likelihood and the prior distribution of the parameters, divided by the posterior distribution and this holds for all parameter values, i.e.:

$$p(\mathbf{X}|K) = \frac{p(\mathbf{X}, \boldsymbol{\theta}_K|K)}{p(\boldsymbol{\theta}_K|\mathbf{X}, K)} \qquad \forall \boldsymbol{\theta}_K \in \boldsymbol{\Theta}_K$$

where $\boldsymbol{\theta}_K$ is a short hand notation for the parameters $\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ under the model with K components. So, by substituting an estimate to $p(\boldsymbol{\theta}_K|\mathbf{X}, K)$ for a suitable chosen $\boldsymbol{\theta}_K$, say $\bar{\boldsymbol{\theta}}_K$, we can estimate the marginal likelihood of M_K, $p(\mathbf{X}|K)$ and so the BF. Chib and Jelaizkov (2001) showed that a suitable estimate of $p(\bar{\boldsymbol{\theta}}_K|\mathbf{X}, K)$ may be obtained on the basis of the Metropolis-Hastings output for sampling $\boldsymbol{\theta}_K$ from its posterior distribution under model M_K; such an algorithm uses as acceptance probability for moving from $\boldsymbol{\theta}_K$ to a proposed $\boldsymbol{\theta}_K^*$

$$\alpha(\boldsymbol{\theta}_K, \boldsymbol{\theta}_K^*) = \min\left\{1, \frac{p(\mathbf{X}, \boldsymbol{\theta}_K^*|K)q(\boldsymbol{\theta}_K|\boldsymbol{\theta}_K^*)}{p(\mathbf{X}, \boldsymbol{\theta}_K|K)q(\boldsymbol{\theta}_K^*|\boldsymbol{\theta}_K)}\right\},$$

where $q(\boldsymbol{\theta}_K^*|\boldsymbol{\theta}_K)$ is the proposal distribution from which $\boldsymbol{\theta}_K^*$ is drawn. In fact, we have

$$p(\bar{\boldsymbol{\theta}}_K|\mathbf{X},K) = \frac{\int_{\Theta_K} \alpha(\boldsymbol{\theta}_K,\bar{\boldsymbol{\theta}}_K)q(\bar{\boldsymbol{\theta}}_K|\boldsymbol{\theta}_K)p(\boldsymbol{\theta}_K|\mathbf{X},K)d\boldsymbol{\theta}_K}{\int_{\Theta_K} \alpha(\bar{\boldsymbol{\theta}}_K,\boldsymbol{\theta}_K)q(\boldsymbol{\theta}_K|\bar{\boldsymbol{\theta}}_K)d\boldsymbol{\theta}_K} = \frac{\mathrm{E}\{\alpha(\boldsymbol{\theta}_K,\bar{\boldsymbol{\theta}}_K)q(\bar{\boldsymbol{\theta}}_K|\boldsymbol{\theta}_K)\}}{\mathrm{E}\{\alpha(\bar{\boldsymbol{\theta}}_K,\boldsymbol{\theta}_K)\}}$$

that, consequently, may be estimated through

$$\hat{p}(\bar{\boldsymbol{\theta}}_K|\mathbf{X},K) = \frac{\sum_{t=1}^{N_1} \alpha(\boldsymbol{\theta}_K^{(t1)},\bar{\boldsymbol{\theta}}_K)q(\bar{\boldsymbol{\theta}}_K|\boldsymbol{\theta}_K^{(t1)})/N_1}{\sum_{t=1}^{N_2} \alpha(\bar{\boldsymbol{\theta}}_K,\boldsymbol{\theta}_K^{(t2)})/N_2},$$

where $\boldsymbol{\theta}_K^{(11)},\ldots,\boldsymbol{\theta}_K^{(N_1,1)}$ is a sample from $p(\boldsymbol{\theta}_K|\mathbf{X},K)$ and $\boldsymbol{\theta}_K^{(12)},\ldots,\boldsymbol{\theta}_K^{(N_2,2)}$ is a sample from $q(\boldsymbol{\theta}_K|\bar{\boldsymbol{\theta}}_K,K)$. Chib and Jelaizkov (2001) also suggested to split the parameters into blocks, which are updated separately (as illustrated in Section 3.1), to increase the estimator efficiency. The point $\bar{\boldsymbol{\theta}}_K$ in practice is chosen as a point of high posterior density, generally the posterior mean of $\boldsymbol{\theta}_K$, in order to maximize the accuracy of the approximation.

5 Application

We show an application of the proposed approach to a real Microarray experiment on a yeast genome (the *Saccharomyces cerevisiae*), aimed at identifying groups of genes involved in the cell cycle and, therefore, characterized by periodic fluctuations in their expression levels. Data refer to $n = 696$ genes observed at $S = 12$ consecutive times during the cell division cycle. A full description of the experiment, carried out by Spellman et al. (1998), and complete data sets are available at http://cellcycle-www.stanford.edu.

The results reported here correspond to 50,000 sweeps of the MCMC algorithm described in Section 3, including a burn-in of 5,000 sweeps. The algorithm seems to mix well over the parameter space and the burn-in seems to be more than adequate to achieve stationarity. This can be seen, for example, in Figure 2(a), which shows the traces of $\boldsymbol{\pi}$ against the number of sweeps (for sake of clarity, data are plotted every 10 sweeps), for the model with $K = 3$ components.

The estimated marginal loglikelihood is plotted in Figure 2(b) against different values of K. It is immediately evident that the model with $K = 3$ components is favored. The BF of this model against the second most favored model, the one with 4 components, is $B_{3,4} = 98716$, implying an overwhelming evidence in favor of the model with $K = 3$, compared to any other model.

The estimated weights we obtained for the 3 groups of genes are respectively $\boldsymbol{\pi} = (0.012, 0.373, 0.615)'$, resulting in a large group including approximately 428 genes, an intermediate group with 244 genes and a residual one made of just 8 genes.

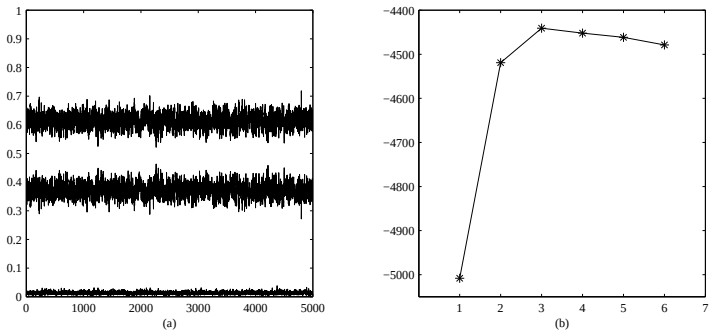

Fig. 2. (a) Traces of $\boldsymbol{\pi}$ against the number of sweeps for the model with three components and (b) marginal loglikelihood for models with up to six components.

Figure 3 shows the estimated mean expression profiles for the three groups of genes. The results we found are in accordance with those obtained by Holter et al. (2000) using a standard value decomposition of the data matrix $\mathbf{X}$. Two dominant periodic patterns, corresponding to the two larger groups, can be recognized. These periodic patterns are out of phase with respect to each other and the maxima and minima in each of them occur at the same time as the maxima and minima in the two main patterns found by Holter et al. (2000).

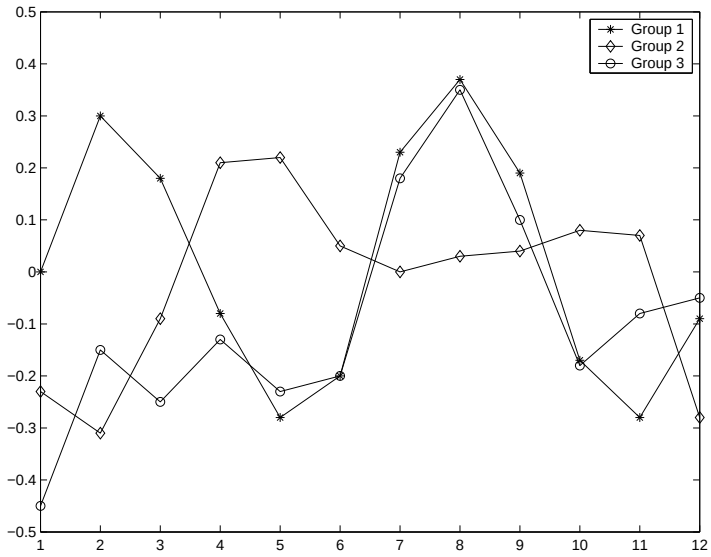

Fig. 3. Mean expression profiles for the three groups.

References

BROWN, M.P.S., GRUNDY, W.N., LIN, D., CRISTIANINI, N., SUGNET, C.W., FUREY, T.S., ARES, M. and HAUSSLER, D. (2000): Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. *Proceeding of the National Academy of Science, 97, 262–267.*

CHIB, S. and JELIAZKOV, I. (2001): Marginal Likelihood from the Metropolis-Hastings Output. *Journal of the American Statistical Association, 96, 270–281.*

FRALEY, C. and RAFTERY, A.E. (1998): How many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis. *The computer journal, 41, 570–588.*

GREEN, P.J. (1995): Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika, 82, 711–732.*

HASTIE, T., TIBSHIRANI, R., EISEN, M.B., BROWN, P.O., ROSS, D., SCHERF, U., WEINSTEIN, J., ALIZADEH, A., STAUDT, L. and BOTSTEIN, D. (2000): Gene Shaving as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns. *Genome Biology, 1, research 0003.*

HOLTER, N.S., MITRA, M., MARITAN, A., CIEPLAK, M., BANAVAR, J.R. and FEDOROFF, N.V. (2000): Fundamental Patterns underlying gene espression profiles: Simplicity from complexity. Proceedings of the National Academy of Sciences, 97, 8409–8414.

JEFFREYS, H. (1939): *The Theory of Probability.* Oxford University Press, Oxford.

KASS, R.E. and RAFTERY, A.E. (1995): Bayes Factors. *Journal of the America Statistical Association, 90, 773–795.*

SPELLMAN, P.T., SHERLOCK, G., ZHANG, M.Q., IYER, V.R., ANDERS, K., EISEN, M.B., BROWN, P.O., BOTSTEIN, D. and FUTCHER, B. (1998): Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell, 9, 3273–3297.*

TANNER, M.A. and WONG, W.H. (1987): The Calculation of Posterior Distributions by Data Augmentation. *Journal of the America Statistical Association, 82, 528–540.*

TAVAZOIE, S., HUGHES, J.D., CAMPBELL, M.J., CHO, R.J. and CHURCH, G.M. (1999): Systematic Determination of Genetic Network Architecture. *Nature Genetics, 22, 281–285.*

YEUNG, K.Y., FRALEY, C., MEURUA, A., RAFTERY, A.E. and RUZZO, W.L. (2001): Model-based Clustering and Data Transformation for Gene Expression Data. *Bioinformatics, 17, 977–987.*