A Stratified Beta-Gaussian Finite Mixture Model for Clustering Genes With Multiple Data Sources

Xiaofeng Dai
Department of Signal Processing,
Tampere University of Techonology
Tampere, Finland
Email: xiaofeng.dai@tut.fi

Harri Lähdesmäki
Department of Signal Processing,
Tampere University of Technology
Tampere, Finland
Helsinki University of Technology
Department of Information and Computer Science
Helsinki, Finland
Email: harri.lahdesmaki@tut.fi

Olli Yli-Harja
Department of Signal Processing,
Tampere University of Techonology
Tampere, Finland
Email: yliharja@cs.tut.fi

***Abstract*—This paper presents a stratified mixture model based clustering framework, sBGMM. It is an extension of one of our previously developed models, BGMM (beta-Gaussian mixture model), which can not only cluster genes based on beta and Gaussian distributed data but also convert information from a third data source to the priors based on which genes are pre-partitioned into several groups. By assigning genes in the same pre-group the same prior probabilities of belonging to a certain cluster, sBGMM transfers information from a third data source into the results and allows a high level of flexibility in the choice of the third data source. Different from data sources that are modeled as the component of the joint model, information used for prior construction can be from any sources and of any level of sparsity. Besides the extremely flexible choice of prior, sBGMM can also be extended to other parametric distributed data, which adds even more flexibility to this model-based clustering framework. We developed an expectation maximization algorithm for jointly estimating the parameters of sBGMM, and propose to tackle model selection problem by approximation based model selection criteria, where four well-known penalized methods, Akaike information criterion, a modified Akaike information criterion, the Bayesian information criterion, and the integrated classification likelihood-Bayesian information criterion, are tested and compared. Both simulation and real case study indicate that information from different data sources can reinforce each other and utilizing information from one data source to stratify the model can improve the clustering accuracy especially when the noise is comparatively high in both beta and Gaussian distributed data. Applications with full set of real mouse gene expression data (modeled as Gaussian distribution) and protein-DNA binding probabilities (modeled as beta distribution) not only yield more biologically reasonable results compared to its non-stratified version, but also discovered the relationship between two set of genes and eight TFs, which are all likely to be involved in Myd88-dependent Toll-like receptor 3/4 (TLR-3/4) signaling cascades.**

***Keywords*—stratified finite mixture model; gene clustering; multiple data fusion; prior**

I. Introduction

Gene clustering has become one of the most explosively expanding tools for genome-level data analysis, such as inferring gene functions [34] and identifying genes involved in a particular molecular pathway [28]. Numerous computational methods have been developed for it, among which the most prevalent ones include hierarchical clustering [9], K-means [15], and Self-Organizing Maps [32]. These approaches are generally applied to gene expression data [16], which although have demonstrated their usefulness in applications [29], are over dependent on the similarity among gene expression patterns, rendering the results less accurate due to the varied transcriptional coherence in response to diverse environmental stresses and vulnerable to system or experimental error because of using single data source alone and no reinforcement from other data sources.

Multiple data fusion has been widely applied to many problems in the field of system biology, assuming that information from different data sources reinforce each other and can offer us a general view of the system from different perspectives. Nowadays, as more and more different biological data sources, such as protein-DNA binding probabilities, protein-protein interactions, evolutionary conservations histone modifications and methylation information, et cetera, are becoming available since new experimental techniques keep emerging, it is possible to cluster genes based on multiple data sources and promising to group genes based on multiple criteria. Therefore, how to efficiently utilize heterogeneous data sources has become one of the most challenging problems in this field.

Gene clustering method can be roughly classified into three categories, which are heuristic, iterative relocation and model-based methods [11]. Common restrictions of using methods that belong to the first two categories are the determination of the number of clusters and handling with the outliers, which however can be easily solved by model-based methods. Due to the clear definition of what a cluster is, a subpopulation with a certain distribution, model-based methods handle with outliers by recasting it as the model selection problem and adding one or more components, respectively [11], [17], [24]. Also, model-based method beats the first two approaches in its statistical nature [11].

In the realm of standard model based gene clustering, besides the most commonly used statistically method, Gaussian mixture model (GMM) [3], [10], [12], [18], [20], [24], [31], [37], mixture models of some other distributions have also

been developed to solve various problems, such as using a two-component beta mixture model (BMM) to cluster correlation coefficients [17] and applying multinomial models to high dimensional text clustering [22], [36]. While most works on model based methods are devoted to exploring novel applications or improving the computational complexity of the algorithm regardless of the information source, [25] proposed a GMM that can incorporate priors beyond expression data by allowing genes that share the same biological function to have an equal prior probability while differ from the other genes in gene clustering.

Inspired by the promising results brought out by stratifying the priors in GMM [25] and the obvious superiority of data fusion over using single data sources alone, we developed a stratified joint mixture model, sBGMM, to cluster genes based on beta and Gaussian distributed data and stratify the prior according to a third data source. This algorithm differs from our previously developed joint mixture model, BGMM [7], in its utilization of three data sources by converting information from a third one to the prior of the joint mixture model. Also, it exceeds the work of [25] by integrating multiple data sources. Moreover, besides the flexible framework inherited from BGMM, sBGMM assigns more freedom to the choice of prior, which is not restricted to any distribution or limited to the completeness of the data.

We have previously developed an approximated (optimize the complete log-likelihood instead of its expectation) expectation maximization (EM) algorithm for BMM, and a hybrid EM algorithm, where EM for beta and Gaussian distributions are approximated and standard version, respectively, for sBGMM. Encouraged by the flexibility provided by sBGMM and its excellent simulation performance shown in [1], we further extend the hybrid EM for sBGMM to the standard EM in this paper, and test it under more simulation scenarios and with real data.

Many statistical methods can be applied to solve the model selection problem, where four well known penalized likelihood criteria (which belong to approximation-based model selection criteria [30]), Akaike information criterion (AIC) [2], [4], modified AIC (AIC3) [4], [5], Bayesian information criterion (BIC) [25], [27], and integrated classification likelihood-BIC (ICL-BIC) [17] are compared in sBGMM in this study. Based on the simulation results, where sBGMM was compared with BGMM under different scenarios, ICL, other than AIC3 which is proposed for being used in the approximated version of sBGMM [1], performs best in sBGMM.

The following sections are organized as 'Methods', 'Results', and 'Conclusions', where mixture model based clustering and EM algorithm are heavily discussed in 'Methods', results of performance test with simulations and real data, as well as a real case application are shown in different subsections of 'Results', and in 'Conclusions' we first summarized this work, and then discussed its limitation and possible extensions.

II. METHODS

This section introduces the proposed algorithm, including the clustering framework, EM algorithm, prior construction, model selection, and its initialization and convergence.

A. Stratified beta-Gaussian mixture model clustering framework

In model-based clustering methods, each observation $\mathbf{x}_j$, where $j = 1, \ldots, n$ and n is the number of genes, is drawn from a finite mixture distribution with the prior probability π_i, component-specific distribution $f_i^{(g)}$ and its parameters θ_i. The formula is given as [21]

$$f(\mathbf{x}_j|\theta) = \sum_{i=1}^{g} \pi_i f_i^{(g)}(\mathbf{x}_j|\theta_i), \tag{1}$$

where $\theta = \{(\pi_i, \theta_i) : i = 1, \ldots, g\}$ is used to denote all the unknown parameters, with the restriction that $0 < \pi_i \leq 1$ for any i and $\sum_{i=1}^{g} \pi_i = 1$. Note that g is the number of components in this model. In the following texts, we ignore the superscript (g) from $f_i^{(g)}$ for simplicity.

In order to integrate as many information sources as possible, we propose in this paper an sBGMM

$$f_{(k)}(\mathbf{x}_j|\theta_{(k)}) = \sum_{i=1}^{g} \pi_{(k),i} f_i^{(g)}(\mathbf{x}_j; \theta_i), \tag{2}$$

where $1 \leq k \leq K$. It means that the genes can be partitioned into several groups, say $G_1, \ldots, G_K$, based on additional prior before EM is run, and the K stratified models share the same set of component distributions while differ in their usage of stratum-specific prior probabilities.

Define $\theta = [\pi, \theta_1, \theta_2]^T$, $\pi = \left[\pi_{(1)}, \ldots, \pi_{(K)}\right]^T$, $\theta_1 = [\alpha_{11}, \ldots, \alpha_{gp_1}, \beta_{11}, \ldots, \beta_{gp_1}]^T$, and $\theta_2 = \left[\mu_{11}, \ldots, \mu_{gp_2}, \sigma_1^2, \ldots, \sigma_{p_2}^2\right]^T$, where p_1 and p_2 each represents the dimension of the observations in BMM and GMM, respectively, and $\pi_{(k)} = [\pi_{(k),1}, \ldots, \pi_{(k),g}]$ where $k = [1, \ldots, K]$ for K stratified models. We also denote Y and Z as the observations of beta distributed and Gaussian distributed data, respectively, function f of $\mathbf{y}$ and f of $\mathbf{z}$ as the density function of beta and Gaussian distribution, respectively, and $\mathbf{x} = [\mathbf{y}^T, \mathbf{z}^T]^T$.

Apart from adding the prior, sBGMM is built from BMM and GMM with the assumption that, for each component i, the beta distributed and Gaussian distributed data are independent. In the BMM part, each component is assumed to be the product of p_1 independent beta distributions, whose probability density function is defined as

$$f_i(\mathbf{y}|\theta_{1i}) = \prod_{u=1}^{p_1} \frac{y_u^{\alpha_{iu}-1}(1-y_u)^{\beta_{iu}-1}}{B(\alpha_{iu}, \beta_{iu})}, \tag{3}$$

where $\theta_{1i} = [\alpha_{i1}, \ldots, \alpha_{ip_1}, \beta_{i1}, \ldots, \beta_{ip_1}]$ and $\mathbf{y} = [y_1, \ldots, y_{p_1}]^T$. Likewise, each component is assumed to follow a Gaussian distribution in the GMM part, whose probability density function of each component for each gene is defined

as

$$f_i(\mathbf{z}|\theta_{2i}) = \frac{1}{(2\pi)^{\frac{p_2}{2}}|V|^{\frac{1}{2}}} \exp\big(-\frac{1}{2}(\mathbf{z}-\mu_i)^T V^{-1}(\mathbf{z}-\mu_i)\big), \quad (4)$$

where $\theta_{2i} = [\mu_{i1},\ldots,\mu_{ip_2},\sigma_1^2,\ldots,\sigma_{p_2}^2]$, $\mu_i = [\mu_{i1},\ldots,\mu_{ip_2}]^T$, $V = \mathrm{diag}(\sigma_1^2,\sigma_2^2,\ldots,\sigma_{p_2}^2)$ and $|V| = \prod_{v=1}^{p_2}\sigma_v^2$. Notice that diagonal covariance matrix is assumed in the Gaussian part, which is especially useful for high-dimensional data since it can significantly reduce the number of parameters that are needed to be estimated from data.

B. EM algorithm

The standard EM algorithm is applied to estimate the parameters θ in sBGMM iteratively, whose derivation is similar with one of our previously developed model, BGMM, as proposed in [6].

The data log-likelihood (natural logarithm is referred to throughout this paper) can be written as

$$\log L(\theta) = \sum_{j=1}^{n}\log\Big(\Big[\sum_{i=1}^{g}\pi_{(k),i}f_i(\mathbf{x}_j|\theta_i)\Big]\Big), \quad (5)$$

given $X = \{\mathbf{x}_j : j = 1,...,n\}$, whose direct maximization, however, is difficult.

In order to make the maximization of Equation 5 tractable, the problem is casted in the framework of incomplete data. Since we assume that the beta and Gaussian distributed data are independent, the complete data likelihood, L_c, can be factored as

$$L_c(\theta) = f(Y|\mathbf{c},\theta)f(Z|\mathbf{c},\theta)f(\mathbf{c}|\theta). \quad (6)$$

If we define $c_j \in \{1,\ldots,g\}$ as the clustering membership of $\mathbf{x}_j$, then the complete data log-likelihood can be written as

$$\log L_c(\theta) = \sum_{j=1}^{n}\sum_{i=1}^{g}\chi(c_j = i)\log\left(\pi_{(k),i}f_i(\mathbf{x}_j|\theta_i)\right), \quad (7)$$

where $\chi(c_j = i)$ is the indicator function of whether $\mathbf{x}_j$ is from the i^{th} component or not.

In the EM algorithm, E step computes the expectation of the complete data log-likelihood

$$\begin{aligned} Q(\theta|\theta^{(m)}) &= E_{\mathbf{c}|X,\theta^{(m)}}(\log L_c) \\ &= \sum_{j=1}^{n} E_{c_j|\mathbf{y}_j,\mathbf{z}_j,\theta^{(m)}}\left[\log\left(f(\mathbf{y}_j|c_j,\theta_1)\right)\right] \\ &\quad + \sum_{j=1}^{n} E_{c_j|\mathbf{y}_j,\mathbf{z}_j,\theta^{(m)}}\left[\log\left(f(\mathbf{z}_j|c_j,\theta_2)\right)\right] \\ &\quad + \sum_{k=1}^{K}\sum_{j\in G_k} E_{c_j|\mathbf{y}_j,\mathbf{z}_j,\theta^{(m)}}\left[\log\big(f(c_j|\pi_{(k)})\big)\right], \end{aligned} \quad (8)$$

where $\theta^{(m)}$ represents the parameters estimated in the m^{th} iteration, and details of the derivation of Q can be found in [21]. By computing the expectation, Equation 8 becomes

$$Q(\theta|\theta^{(m)}) = \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log(\pi_{(k),i}f_i(\mathbf{y}_j|\theta_{1i})f_i(\mathbf{z}_j|\theta_{2i})), \quad (9)$$

where

$$\begin{aligned} \tau_{ji}^{(m)} &= p(c_j = i|\mathbf{x}_j,\theta^{(m)}) \\ &= \frac{\pi_{(k),i}^{(m)}f_i(\mathbf{y}_j|\theta_{1i}^{(m)})f_i(\mathbf{z}_j|\theta_{2i}^{(m)})}{\sum_{i'=1}^{g}\pi_{(k),i'}^{(m)}f_{i'}(\mathbf{y}_j|\theta_{1i'}^{(m)})f_{i'}(\mathbf{z}_j|\theta_{2i'}^{(m)})}, \end{aligned} \quad (10)$$

is the estimated posterior probability of $\mathbf{x}_j$, which belongs to the k^{th} layer according to the prior, coming from component i at iteration m according to Bayes' rule. Note that we can assign each $\mathbf{x}_j$ to the component i_0 that maximizes its estimated posterior probability, i.e., $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$. Also, the assumption that the beta distributed and Gaussian distributed data are independent is carried over to the expected log-likelihood as shown by Equations 8 and 9.

To derive the closed form or numerical optimization formula for updating parameters in sBGMM, we used Lagrange multipliers to solve this constrained optimization problem, with the Lagrangian function shown in Equation 11.

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log\left(f_i(\mathbf{y}_j|\theta_{1i})\right) \\ &\quad + \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log\left(f_i(\mathbf{z}_j|\theta_{2i})\right) \\ &\quad + \sum_{k=1}^{K}\sum_{j\in G_k}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log(\pi_{(k),i}) \\ &\quad + \sum_{k=1}^{K}\lambda_k\left(1 - \sum_{i'=1}^{g}\pi_{(k),i'}\right) \end{aligned} \quad (11)$$

Parameters of BMM part, $\theta_{1i} = [\alpha_{i1},\ldots,\alpha_{ip_1},\beta_{i1},\ldots,\beta_{ip_1}]$ $1 \le i \le g$, are optimized by Newton-Raphson method and updated by

$$\theta_{1i}^{(m+1)} = \theta_{1i}^{(m)} - H^{-1}(\theta_{1i}^{(m)})\nabla_{\theta_{1i}}\mathcal{L}(\theta_{1i}^{(m)}), \quad \theta_{1i} \ge \mathbf{1}, \quad (12)$$

where $H^{-1}(\theta_{1i}^{(m)})$ is the inverse of the Hessian matrix evaluated at $\theta_{1i}^{(m)}$, and $\mathcal{L}(\theta_{1i}^{(m)})$ is the Lagrangian function of $Q(\theta_{1i}^{(m)})$.

Parameters of the GMM part, $\theta_{2i} = [\mu_{i1},\ldots,\mu_{ip_2},\sigma_1^2,\ldots,\sigma_{p_2}^2]$ $1 \le i \le g$, in sBGMM can be estimated by the standard EM algorithm of GMM with diagonal covariance matrix as shown in the following closed form formula

$$\hat{\mu}_{iv}^{(m+1)} = \sum_{j=1}^{n}\tau_{ji}^{(m)}z_{jv}/\sum_{j=1}^{n}\tau_{ji}^{(m)}, \quad (13)$$

$$\hat{\sigma}_v^{2,(m+1)} = \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ji}^{(m)}(z_{jv} - \mu_{iv}^{(m)})^2/n, \quad (14)$$

which can be obtained by plugging Equation 4 in Equation 11

and taking the derivatives of Equation 11 with respect to μ_{iv} and σ_v^2, respectively.

Optimization of the prior probability of each gene's clustering membership, π, can be derived by taking the derivative of Equation 11 with respect to $\pi_{(k),i}$, i.e.,

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \pi_{(k),i}} &= \sum_{j \in G_k} \tau_{ji}^{(m)} \frac{1}{\pi_{(k),i}} - \lambda_k, \\ \hat{\pi}_{(k),i}^{(m+1)} &= \sum_{j \in G_k} \tau_{ji}^{(m)} / \lambda_k.\end{aligned}$$

Moreover, since

$$\begin{aligned}1 &= \sum_{i=1}^{g} \pi_{(k),i} \\ &= \sum_{i=1}^{g} \frac{1}{\lambda_k} \sum_{j \in G_k} \tau_{ji} \\ &= \frac{1}{\lambda_k} \sum_{j \in G_k} \sum_{i=1}^{g} \tau_{ji} \\ &= \frac{1}{\lambda_k} \sum_{j \in G_k} 1,\end{aligned}$$

thus, $\lambda_k = n_k$. Consequently, the updates of π is given by

$$\hat{\pi}_{(k),i}^{(m+1)} = \sum_{j \in G_k} \tau_{ji}^{(m)} / n_k, \tag{15}$$

where G_k is the k^{th} group with n_k genes, according to the prior.

From the above equations, it is easy to see that the EM of sBGMM will reduce to the EM of BGMM if $K = 1$, and will further reduce to BMM or GMM, respectively, as p_2 or p_1 equals to zero.

1) Prior construction: Priors of Equation 2 can be determined from any possible data sources. It can be either another complete data source different from what have been used in the component models (BMM and GMM), e.g., the pre-cluster results obtained from PPI data [35], or some incomplete information relevant to our problem, e.g., information retrieved from database. In the following study, we employ a complete PPI data set for simulation test, and obtain a set of incomplete information from a database for real case study. Conversion of PPI data into prior is described below.

PPI data, which is typically a binary square matrix, is first converted into contact matrix (denoted as A) and then transformed into pathlength matrix (denoted as P). Contact matrix is in the form of

$$A = \begin{cases} 1 & \text{if } i \Leftrightarrow j \\ 0 & \text{if } i \not\Leftrightarrow j, \end{cases} \tag{16}$$

where $i \Leftrightarrow j$ means the existence of a connection between node i and j while $i \not\Leftrightarrow j$ denotes the other way around. In the pathlength matrix, the pathlength between nodes i and j is denoted as P_{ij} and characterized as the smallest integer $k \geq 1$ such that $(A^k)_{ij} \neq 0$. P contains all the path lengths for all pairs of nodes which are calculated by the 'pathlength' function of the 'CONTEST' toolbox in matlab [33]. We use the pathlength matrix to pre-cluster the genes (corresponding to the proteins they encode) using a simple hierarchical clustering algorithm which employs Euclidean distance as the distance matrix and nearest neighbor algorithm as the linkage construction method, and matlab function 'clusterdata' is used here for this purpose. Then we assume that genes from the same pre-cluster share the same prior probability $\pi_{(k),i}$ of coming from the same cluster i, and allow them coming from different clusters.

C. Model Selection

Four well-known approximation-based model selection criteria, BIC [25], [27], ICL [17], AIC [2], [4], and AIC3 [4], [5] are compared in sBGMM, according to which the best-performing criterion within the tested scope is chosen. Calculations for the above criteria are defined as

$$AIC = -2 \log L(\hat{\theta}) + 2d, \tag{17}$$

$$AIC3 = -2 \log L(\hat{\theta}) + 3d, \tag{18}$$

$$BIC = -2 \log L(\hat{\theta}) + d \log(nM), \tag{19}$$

$$ICL = -2 \log L(\hat{\theta}) + d \log(nM) - 2 \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ji} \log(\tau_{ji}), \tag{20}$$

where d is the number of free parameters, and M (in equations 19 and 20) is the total amount of the data ($M = \sum_{w=1}^{W} M_w$, M_w is the size of data set w and W is the number of input data sets). Note that $-2 \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ji} \log(\tau_{ji})$ is the estimated entropy of the fuzzy classification matrix $C_{ji} = (\tau_{ji})$ [17].

sBGMM has $K - 1$ more free π_i's than BGMM because of the K stratified layers. In BGMM, the number of free parameters d is the summation of those in BMM and GMM minus one set of redundant free π_i's, which is $d_{BG} = 2gp_1 + p_2 + p_2 g + (g - 1)$. Therefore, the number of free parameters in sBGMM is $d_{sBG} = 2gp_1 + p_2 + p_2 g + K(g-1)$.

D. Initialization and convergence

In this study, parameters α_{iu}'s and β_{iu}'s for each dimension of beta distribution u ($u \in \{1, \ldots, p_1\}$) are initialized by method-of-moments so that their means are randomly distributed within the range of $y_{1u}, \ldots, y_{nu}$ and variances are equal for all clusters (g), μ_{iv}'s and σ_v^2's are obtained from the randomly initialized fuzzy c-means clustering results, and π_i's are initialized with the same random value within each group G_k, and the sum of the probabilities of g components is one.

In order to avoid the possible local maxima, we run the algorithm multiple (100) times with different initial values. The convergence threshold (where Q is used to monitor the convergence) and maximum number of iterations were set to 0.0001 and 100, respectively, for all the tested models, and all the simulations have reached their convergence according to the statistics stored during the simulations.

III. RESULTS

We first tested the performance of sBGMM with artificial and real data, respectively, and then applied it to a real biological case, which are discussed separately below.

A. Performance test with artificial data

According to work done in [19] only part of protein-DNA binding data and gene expression data agree with each other (consisting of the same number of clusters), and data can fall into three regions as illustrated in Figure 1. Beta and Gaussian distributed data may share the same number of underlying clusters (denoted as 'Region 1') or may not, and we denote the scenario that beta distributed data has more underlying clusters as 'Region2', and 'Region3' for the scenario of the other way around. To match the three scenarios, we designed data sets 1 to 3 for data of both beta and Gaussian distributions, respectively, whose parameters are listed in Table I. Each artificial data set is designed to fall into five categories: 'good Beta' (gB), 'bad Beta' (bB), 'good Gaussian' (gG), 'bad Gaussian due to close means' (bG_m), and 'bad Gaussian due to large variances' (bG_v), where 'good' stands for low noise level, and 'bad' means the opposite. The dimensions are designed to be $n = 100$ and $p = 4$ for both data sets. We also designed three PPI data sets to test the influence of different priors on the clustering results. Prior 1 and 2 are constructed based on the same underlying ground truth (the same number of underlying clusters and the same clustering membership of each gene) but differ in their noise levels, while prior 3 contains some mis-clustering (clustering membership of some genes are not consistent with the designed Gaussian and beta distributed data) information and shares the same noise level with Prior 1. All sparsity patterns are shown in Figure 2, where the three priors are denoted as 'T9', 'T2', and 'F9', respectively, with the capital letter representing 'true' or 'false' (meaning that there is or there is not mis-clustering information, respectively), and the following number standing for the noise level (the higher the number the lower the noise), e.g., 9 means the intensity of signal over background is 9. All the simulations are repeated 10 times with randomly generated data sets (including the data used for prior construction).

We used the same scoring system as developed in [7] for performance evaluation, which is denoted as 'E score'

$$\begin{aligned} e_j(r) &= \begin{cases} 1 & if \;\; \hat{z}_{ji} = 1 \;\; and \;\; r_i = T_j \\ 0 & otherwise \end{cases} \\ E &= \max_{r \in R} \sum_{j=1}^{n} e_j(r)/n \qquad (21) \\ R &= \big\{ r = (r_1, \ldots, r_{\hat{g}}) : \forall i \neq j \; r_i \neq r_j; \\ &\quad r_i \in \{1, \ldots, \max\{\hat{g}, g\}\} \big\}. \end{aligned}$$

Notations of in this scoring system are defined as follows. T_j denotes the ground truth clustering membership of data j. R stands for all possible associating ways between the estimated and the true clusters, where r_i is the label of data belonging to component i predicted by the clustering algorithm, and r is chosen from labels $1, 2, \ldots, \max\{\hat{g}, g\}$ ($\hat{g}$ and g are the largest labels in the estimated and ground truth clustering respectively). Denote also e as the individual score of each gene, E as the average score of all the genes for each repetition, 'E score' of each repetition as the one corresponding to the optimal Q, and the final 'E score' of each data set as the median of the 20 'E score's. This scoring system evaluates the overall performance of the model since it not only records the accuracy of the results but also reflects the influence of the criterion for model selection.

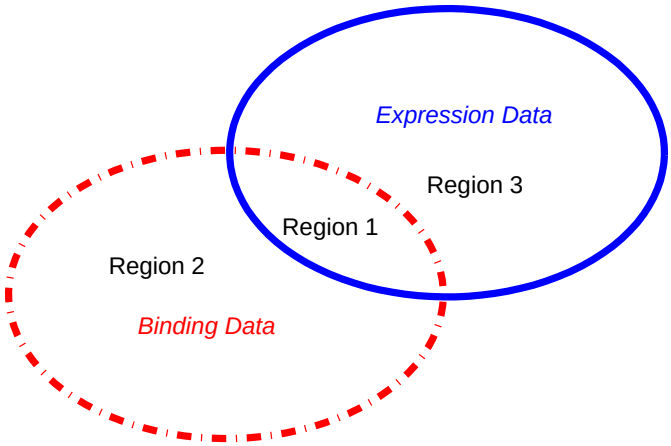


Figure 1. Region divisions of input data. In Region 1 gene expression and protein-DNA binding data have the same number of underlying components; in Region 2 binding data has more components; in Region 3 expression data has more components.

We compared the performance of sBGMM and its non-stratified version, BGMM, with data set 1 to data set 3, each coupled with prior 'T9', 'T2' and 'F9'. Before performance test, we first compared each model selection criterion in handling different scenarios in each model, whose results are shown in Table II. According to the average E scores shown in Table II, there is no universal optimal criterion for sBGMM or BGMM, but ICL is much safer to choose for sBGMM since it selects most of the correct models.

Performance comparison results of sBGMM with its non-stratified version under different scenarios with different priors are shown in Figure 3, where the E scores are calculated with the assumption that the real number of underlying clusters is three (therefore the prior is designed to contain three underlying clusters) and the model is chosen by the criterion that generates the highest average E score under each scenario. It is seen from Figure 3 that sBGMM and BGMM perform equally well when at least one type of data (excluding the prior) contains less noise and has the correct number of underlying clusters for data within 'Region 1', anything combined with 'gB' for data within 'Region 2', and anything combined with 'gG' for data within 'Region 3'. This indicates that our joint models (both sBGMM and BGMM) have the ability to offset the noisy or incorrect information within one type of data by utilizing information from the other one. However, when the noise level, including noise and

Figure 2. Sparsity patterns of the contact matrix of the artificial PPI data sets. (a) 'T9': true prior with noise level equals 9. (b) 'T2': true prior with noise level equals 2. (c) 'F9': false prior with noise level equals 9.

incorrect number of underlying clusters, is too high for both types of data, using additional information becomes important as shown by 'yellow' and 'carmine' in 'Region 1', 'blue', 'yellow' and 'carmine' in 'Region 2', and 'cyan', 'yellow', 'red' and 'carmine' in 'Region 3'. It is also clear that there is no significant difference for using different priors ('T9', 'T2', and 'F9') in sBGMM if the prior does not contain too much mis-clustering information, which means that sBGMM is not sensitive to the noise and is tolerant of small amount of inconsistent information in the prior. All together, these results indicate that adding additional prior can utilize information from more data sources, rendering sBGMM more robust in handling with various scenarios than BGMM.

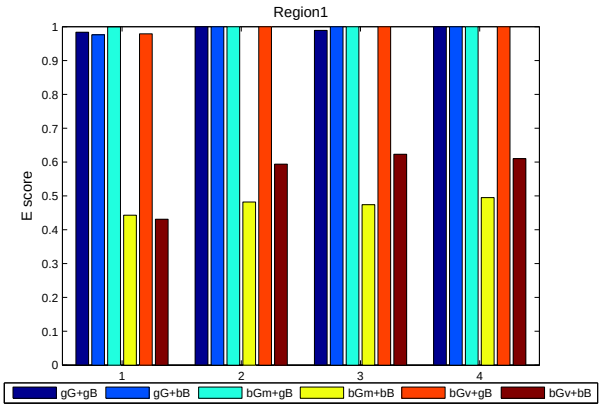
M	P	R	AIC	AIC3	BIC	ICL
BGMM		R1	0.8270	0.8325	0.8459	**0.8464**
		R2	**0.7855**	0.7796	0.7663	0.7723
		R3	0.7763	0.7803	**0.7910**	0.7849
sBGMM	T9	R1	0.8545	0.8680	0.8806	**0.8845**
	T9	R2	0.7434	0.7714	0.8376	**0.8430**
	T9	R3	0.7820	0.7995	0.8188	**0.8270**
sBGMM	T2	R1	0.8459	0.8636	**0.8844**	0.8820
	T2	R2	0.7653	0.8001	0.8305	**0.8391**
	T2	R3	0.7699	0.7941	0.8323	**0.8343**
sBGMM	F9	R1	0.8444	0.8600	**0.8881**	**0.8881**
	F9	R2	0.7430	0.7674	0.8406	**0.8430**
	F9	R3	0.7575	0.7900	**0.8295**	**0.8295**

Note: Values shown here are the averages of E scores over all the tested cases ('gG+gB', 'gG+bB', 'bG$_m$+gB', 'bG$_m$+bB', 'bG$_v$+gB', 'bG$_v$+bB') selected by each criterion in each model. 'P' column shows the priors. 'M' column lists the name of the tested models. 'R' column shows the region that beta and Gaussian distributed data belong to. 'ICL' is short for 'ICL-BIC'. E scores shown in bold face are the selected best criterion with respect to highest average E scores and used in drawing Figure 3. All values are rounded to four decimal points.

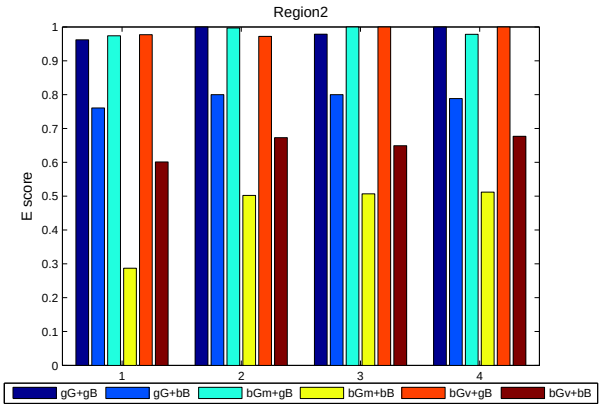
Table II
COMPARISON OF DIFFERENT MODEL SELECTION CRITERIA IN SBGMM AND BGMM.

B. Performance test with real data

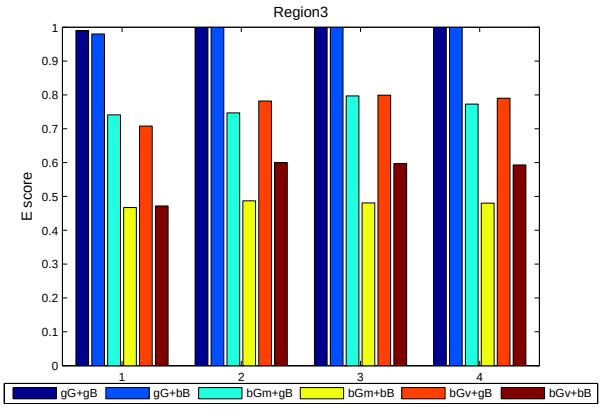
We applied our methods to mouse protein-DNA binding probabilities (modeled as beta distribution) and gene expression data (modeled as Gaussian distribution). The protein-DNA binding data contains the probabilities of 266 TFs binding to 20397 genes, which were calculated with mouse-specific position weight matrices from the TRANSFAC database (the web server and data are available at http://xerad.systemsbiology.net/ProbTF/ [14]). The gene expression data is composed of 1960 genes measured from 95 conditions [26], where six Toll-like receptor (TLR) agonists (C_pG, Pam$_2$CSK$_4$, Pam$_3$CSK$_4$, LPS, poly I:C and R848) were used as the treatments, and four gene knock-out mutants and different time points were included to increase the diversity of the TLR-stimulated gene expression data set and the number of measurements. There are 1766 genes measured in both datasets. We removed the genes whose gene expression profiles have low absolute values (less than 10th percentile) with matlab function 'genelowvalfilter', and then chose genes whose annotations are available through the functional classification tool of DAVID database (whose web server is available at http://david.abcc.ncifcrf.gov/home.jsp [13]). In the end, 673 genes are chosen for the following studies. The chosen protein-DNA binding data (beta distributed data that are used in this study) is composed of the binding probabilities of the 673



(a)



(b)



(c)

Figure 3. Simulation results. Performance comparison of sBGMM with BGMM for (a) region 1 data, (b) region 2 data, and (c) region 3 data. In x-axis of each region, '1' to '4' represent different models, each corresponds to BGMM, sBGMM ('T9'), sBGMM ('T2'), sBGMM ('F9'), accordingly. The y-axis represent the E scores.

genes for six TFs ('Junb', 'Jund1', 'Jun', 'Fos', 'Fosb', and 'Cebpb') which are involved in AP1 gene regulatory network in mouse according to TRED (Transcriptional Regulatory Element Database, which is available at http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home). The Gaussian distributed data that are fitted in the model are the gene expression data of these 673 genes at 23 conditions, which are the midpoint of each time series (selected time point for each treatment are shown in Table III).

Point	Treatment	Time
1	Atf_3^-	0
2	C_pG+Atf_3^-	120
3	LPS+Atf_3^-	240
4	Pam_2CSK_4+Atf_3^-	120
5	poly I:C+Atf_3^-	120
6	$Crem^-$	0
7	LPS+$Crem^-$	240
8	poly I:C+$Crem^-$	360
9	$Myd88^-$	0
10	LPS+$Myd88^-$	60
11	Pam_3CSK_4+$Myd88^-$	60
12	poly I:C+$Myd88^-$	60
13	$TicamI^-$	0
14	LPS+$TicamI^-$	120
15	LPS+Pam_2CSK_4+$TicamI^-$	120
16	no	0
17	C_pG	60
18	LPS	360
19	Pam_2CSK_4	80
20	Pam_3CSK_4	240
21	Pam_3CSK_4+poly I:C	60
22	poly I:C	120
23	R848	120

Note: 'Point' refers to the labels of the x axis; '-' means the mutant strain that does not have the particular gene; time point chosen for the treatment is shown in the column 'Time', and time unit is 'min'. Treatments after point 16 were all applied to the wild type.

Table III
TREATMENTS OF THE GENE EXPRESSION DATA

We constructed two types of priors for real case performance test. Type 1 prior contains two priors (denoted as '$P1_a$', '$P1_b$') which both utilize the transcriptional regulation information stored in TRED. Curation of transcriptional regulation in TRED are done with both experimental evidence and promoter finding tools, and currently involves genes within 36 cancer-related TF families. '$P1_a$' is built from the clustering information of AP1 network, where 10 genes are assigned to two clusters, and the memberships of the rest genes are left unspecified. '$P1_b$' keeps all the clustering membership in '$P1_a$' intact, and specifies the rest memberships from all the other networks in TRED by removing genes that are involved in several networks or form singleton clusters (resulting in 16 more memberships specified). Type 2 priors are obtained from an online classification tool DAVID (the web server is available at http://david.abcc.ncifcrf.gov/gene2gene.jsp [13]), a

database for annotation, visualization and integrated discovery. Different thresholds ('Highest', 'Medium', 'Lowest') were set to obtain different functional classification results, resulting in three different type 2 priors, which are denoted as '$P2_a$', '$P2_b$', '$P2_c$', respectively.

We tested the performance of sBGMM by comparing its performance on clustering the 673 genes (with both types 1 and type 2 priors) with its non-stratified form (BGMM), and both of its component models (BMM, GMM). We employed Gene Ontology (GO) in this study to validate the clustering results. In order to find the most significant annotated terms by looking at the probabilities that the terms are counted by chance, we used the hypergeometric probability distribution to calculate the p-values of gene enrichment score (called 'p-values' for simplicity) for each cluster by each model with each model selection criterion (Bioinformatics Toolbox 3.1 in Matlab). We compared the means and medians of each clustering results by each model with respect to different aspects (molecular function, cellular component, biological process, and all aspects, which are denoted as 'F', 'C', 'P' and 'All'), whose results are shown in Table IV. To see how stable the algorithms work with our test data set, or in other words, whether 100 iterations are enough for convergence, we repeated each set of iterations (100) three times for different models, with one repetition for each model shown in Table IV.

There are at least four pieces of information unveiled by Table IV. First, BGMM works better than BMM and GMM with respect to smaller means and medians of the group p-values. Second, sBGMM can significantly improve the clustering performance compared with BGMM and its component models when the prior is properly chosen. As shown in this table, results generated with type 1 priors are better than those with type 2 priors, whose means and medians are significantly smaller than those of BGMM, BMM and GMM; moreover, sBGMM with type 1 priors generate more stable results than the other models, i.e., two out of three repetitions of sBG_{P1_a} and all three repetitions of sBG_{P1_b} converge to the same clustering, respectively. This is because information delivered by type 1 priors are consistent with that used for choosing TFs of protein-DNA binding probabilities, while type 2 priors, which are the classification results from an online functional classification tool, might group the same gene into another cluster based on its own criteria (DAVID groups genes by measuring the functional relationship of gene pairs based on the similarity of their global annotation profiles [13]). Third, $P1_b$ is denser than $P1_a$, and generates more stable results (three vs. two repetitions converge to the same result), which indicates that the more consistent (consistent with data) information carried out by the prior the more accurate the results will be. However, both type 1 priors used here are quite sparse, therefore, we expect to get even higher accuracy if denser and consistent (consistent with data) prior is available. Fourth, sBGMM with type 1 prior works better than DAVID functional classification tool. As shown in Table IV, all the evaluated quantities of the results obtained from DAVID ($P2_a$, $P2_b$, $P2_c$) are worse than those of sBGMM (coupled with type 1 priors), BGMM, and even some of the results of GMM. Although the improved accuracy can not show the superiority of sBGMM over DAVID since totally different types of data sources are used, the results demonstrate the power of employing gene expression and protein-DNA binding data in gene clustering over relying on global annotation profiles.

C. Biological application with sBGMM

After performance test, we further analyzed all the 1766 genes in our data set. We compared the 1766 genes with the genes involved in all the 36 cancer related gene networks stored in TRED, and decided to extract the information from the network that has the largest overlap with our gene set (NFKB network) for further analysis. There are seven TFs involved in this network, out of which five (which are 'Rel', 'Nfkb1', 'Msx1', 'Rela', 'Myb', and named TF_{normal} for convenience) are available in our data set. Protein-DNA binding probabilities of those five TFs to all the 1766 genes and gene expression data of the 23 midpoint conditions (midpoint of each time series) were chosen as the beta distributed and Gaussian distributed data set, respectively. Genes involved in NFKB network are grouped into six clusters by TRED, among which 42 are present in our data set. We constructed a type '$P1_b$' prior for the whole gene set since it tends to have a more stable behavior compared with '$P1_a$' according to the real case performance test (see the previous subsection).

There are 34 genes that encode TFs (named TF genes) among the whole data set. We first clustered the 34 TF genes with BGMM, for three times, and chose the TF gene cluster which has the smallest enrichment p-values for further analysis. There are 11 genes in the selected TF gene group, out of which eight are repeated clustered together among three repetitions and, for convenience, we call them the 'core TF genes' and denoted as TF_{core} in the following text.

To find the influence of the choice of protein-DNA binding probabilities on the clustering accuracy of sBGMM and find a set of protein-DNA binding probabilities as suitable as possible for further analysis, we first clustered the 1766 genes by sBGMM with binding data corresponding to TF_{normal}, and then re-clustered them with those selected by TF_{core}, each with three repetitions. For comparison purpose, we also clustered the 1766 genes by BGMM with binding data of the core TF genes, and the result of one repetition from each clustering were compared and shown in Table V. Note that the expression data and the prior are the same in the models where they were used.

It is interesting to see from Table V that the group p-values are significantly dropped after using the core TF genes for gene clustering, and the group p-values obtained with sBGMM are overwhelmingly lower than those obtained by BGMM. This means that the core TF genes are more responsible to TLR-stimulated macrophage activation than the TFs chosen based on the prior information obtained from NFKB network, and again demonstrates the superiority of sBGMM over its non-stratified version.

We further analyzed the causal relationship between the set of core TFs and the whole set of genes. We notice that among the eight core TF genes, 'E2F6', 'E2f7', 'Foxm1' and 'Nfatc1' are clustered together with 363 other genes, and 'Rest', 'Rfx5', 'Mxd1' and 'Stat1' fall into the same group with 305 other genes. Moreover, by examining the expression profiles of the two sets of genes under different treatment (shown in Figure 4), it is clear that there is a plateau existed in all profiles from point 26 and 48 where either mutant *Myd88*$^-$ or *TicamI*$^-$ is used, or no treatment is applied or C_pG is added. This indicates that genes *Myd88* and *Ticam1* are crucial for the system (which involves the genes that belong to the four clusters) to response to the external stimuli, and agonist C_pG does not have so much influence on it. Moreover, whenever LPS or poly I:C is added to the wild type (regions between points 5 and 10, 14 and 16, 18 and 22, 23 and 26, 48 and 59, 79 and 87), there is a sharp drop in the red profile while there is a peak in the green curve. This feature indicates that the two set of genes (including the core TF genes) are sensitive to LPS and poly I:C, and behave in an opposite way after being stimulated. Genes that are clustered with 'E2F6', 'E2f7', 'Foxm1' and 'Nfatc1' are activated by them while repressed by TFs 'Rest', 'Rfx5', 'Mxd1' and 'Stat1'; while operation goes the other way around for the other set of genes. Moreover, since poly I:C, LPS and C_pG are TLR-3, TLR-4 and TLR-9 agonists, respectively, and Myd88 and Ticam1 are adaptors involved in TLR-3/4 signaling according to [23], we can deduce that most of the two set of genes (including TF genes) are involved in Myd88-dependent TLR-3/4 signaling cascades.

IV. Conclusion and Future Work

This paper presents a novel method based on stratified beta-Gaussian mixture model, sBGMM, for gene clustering from multiple data sources. In addition to integrating beta distributed and Gaussian distributed data, sBGMM can also facilitate clustering by employing priors which come from a third data source. A stratified version of EM algorithm is developed for jointly estimating parameters from beta and Gaussian distributions, and is used as the core of sBGMM. sBGMM differs from its non-stratified version (BGMM) by setting the same prior probabilities of coming from each cluster to genes that belong to the same layer which are stratified according to the additional prior. In principle, any relevant information can be used as priors, whereas in this study, we built the prior from PPI data in simulations, and retrieved it from database TRED in the real case study. Simulation results show that sBGMM works better than its non-stratified version especially when both beta and Gaussian distributed data contain too much noise, and certain mis-clustering information in the prior is tolerable. In real case study we not only demonstrated the superiority of sBGMM compared with BGMM and a gene annotation based classification method (DAVID functional classification tool) by analyzing 673 genes, but also revealed the relationship of two sets of genes and eight TFs in TLR-stimulated macrophage signaling through analyzing the full data set (1766 genes).

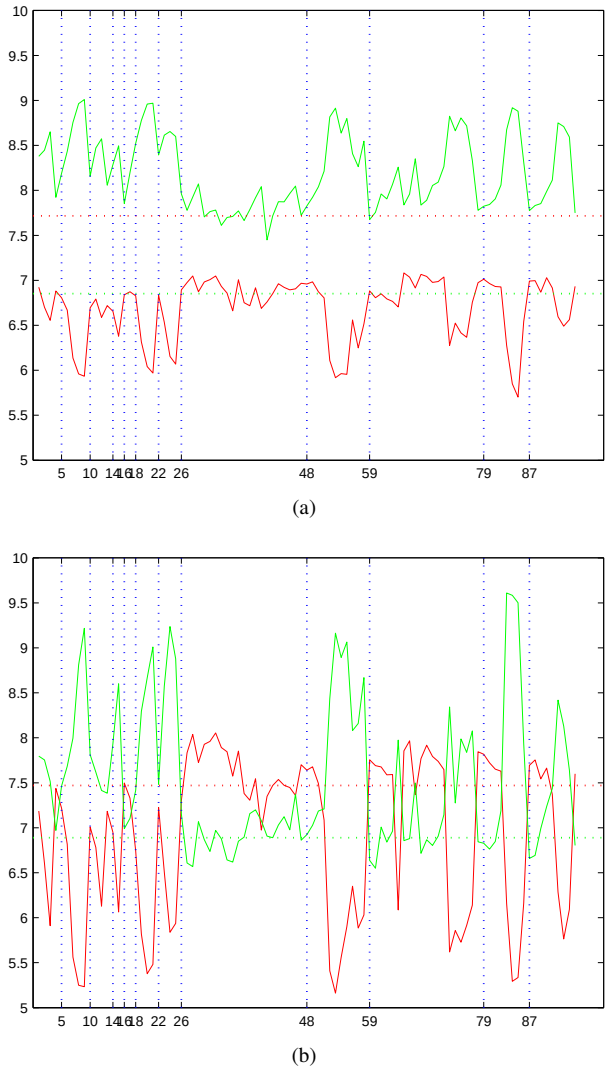


Figure 4. Median gene expression profiles of the (a) interested genes and (b) TFs. Solid curves represent the median expression profile of the genes or TFs. Horizontal dot lines stand for the expression level of wild type without treatment. Vertical blue lines divides the whole plane into different regions, where in each region different treatment is applied.

This work demonstrates one approach of utilizing multiple data sources in gene clustering, and data of other distributions can also be incorporated into this framework by joining EM algorithm of that particular distribution in a similar way. So in a sense, the framework proposed in this paper is applicable to many problems and not limited to the particular problem considered here [8].

Although, sBGMM is tolerant to some mis-clustering information in the prior, its performance might not be improved or even dragged down if the prior is built under different criterion and totally irrelevant to the focused problem. Moreover, although sBGMM is extremely useful when only sparse prior is available, it might not be able to efficiently utilize the third data source whose information is complete (such as PPI data). Moreover, since PPI data is one direct

measure of the regulatory network and is commonly used in gene clustering for many applications, such as inferring gene functions [34] and discovering genes involved in a particular molecular pathway [28], it is important to develop a model that can make as efficient use of PPI data as possible. In the future, instead of utilizing PPI data as prior, we could model it as Bernoulli distribution and treat it as one component of the joint model-based clustering framework.

ACKNOWLEDGMENT

This work was supported by the Academy of Finland (application number 129657, Finnish Programme for Center of Excellence in Research 2006-2011). We would also like to thank the Tampere Graduate School in Information Science and Engineering (TISE) for its financial support in this project.

REFERENCES

[1] X. F. Dai, H. Lähdesmaki, and O. Yli-Harja, *sBGMM: a stratified Beta-Gaussian mixture model for clustering genes with multiple data sources*. International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies (BIOTECHNO 2008), Bucharest, Romania, 29 June - 5 July 2008, pp. 94-99.

[2] H. Akaike, *A new look at the statistical identification model*. IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716-723, 1974.

[3] J. D. Banfield and A. E. Raftery, *Model-based Gaussian and non-Gaussian clustering*. Biometrics, vol. 49, no. 3, pp. 803-821, 1993.

[4] C. Biernacki and G. Govaert, *Choosing models in model-based clustering and discriminant analysis*. J. Statis. Comput. Simul., vol. 64, pp. 49-71, 1999.

[5] H. Bozdogan, *Model Selection and Akaike Information Criterion (AIC): The General Theory and its Analytic Extensions*. Psychometrika, vol. 52, pp. 345-370, 1987.

[6] X. F. Dai, T. Erkkilä, O. Yli-Harja, and H. Lähdesmaki, *A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data*. BMC Bioinformatics, accepted.

[7] X. F. Dai, H. Lähdesmaki, and O. Yli-Harja, *BGMM: a Beta-Gaussian mixture model for clustering genes with multiple data sources*. Fifth international workshop on computational system biology (WCSB 2008), Leipzig, Germany, 11 - 13 June 2008, pp. 25-28.

[8] X. F. Dai, O. Yli-Harja, and A. S. Ribeiro, *Determing noisy attractors of delayed stochastic Gene Regulatory Networks from multiple data sources*. submitted.

[9] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*. Proceedings of the National Academy of Sciences of the United States of America, vol. 95, pp. 14863-14868, 1998.

[10] C. Fraley, *Algorithms for model-based Gaussian hierarchical clustering*, SIAM Journal on Scientific Computing, vol. 20, no. 1, pp. 270-281, 1999.

[11] C. Fraley and A. E. Raftery, *Model-based clustering, discriminant analysis, and density estimation*, Journal of the American Statistical Association, vol. 97, no. 458, pp. 611-631, 2002.

[12] D. Ghosh and A. M. Chinnaiyan, *Mixture modeling of gene expression data from microarray experiments*. Bioinformatics, vol. 18, no. 2, pp. 275-286, 2002.

[13] G. D. Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, *DAVID: Database for Annotation, Visualization, and Integrated Discovery*, Genome Biology, vol. 4, no. 9, pp. R60, 2003.

[14] H. Lähdesmäki, A. G. Rust, and I. Shmulevich, *Probabilistic Inference of Transcription Factor Binding from Multiple Data Sources*, PLoS ONE, vol. 3, no. 3, pp. e1820, 2008.

[15] R. Herwig, A. J. Poustka, C. Muller, C. Bull, H. Lehrach, and J. O'Brien *Large-scale clustering of cDNA-fingerprinting data*, Genome Research, vol. 9, no. 11, pp. 1093-1105, 1999.

[16] D. X. Jiang, C. Tang, and A. D. Zhang, *Cluster analysis for gene expression data: a survey*, IEEE Transactions on knowledge and data engineering, vol. 16, no. 11, pp. 1370-1386, 2004.

[17] Y. Ji, C. Wu, P. Liu, J. Wang, R. K. Coombes, *Applications of beta-mixture models in bioinformatics*. Bioinformatics, vol. 21, no. 9, pp. 2118-2122, 2005.

[18] H. Li and F. Hong *Cluster-rasch models for microarray gene expression data*, Genome Biology, vol. 2, no. 21, pp. research0031.1-0031.13, 2001.

[19] M. J. Herrgard, B. Lee, V. Portnoy, and B. Palsson, *Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces*, Genome Research, vol. 16, pp. 627-635, 2006.

[20] G. Mclachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.

[21] G. Mclachlan and D. Peel, *Finite mixture models*, John Wiley & Sons, Manhattan, USA, 2000.

[22] M. Meila and D. Heckerman, *An experimental comparison of model-based clustering methods*, Machine Learning, vol. 42, pp. 9-29, 2001.

[23] L. A. O'Neill, K. A. Fitzgerald, and A. G. Bowie, *The Toll-IL-1 receptor adaptor family grows to five members*, Trends in Immunology, vol. 24, no. 6, pp. 286-290, 2003.

[24] W. Pan, J. Z. Lin, and C. T Le, *Model-based cluster analysis of gene expression data*. Genome Biology, vol. 3, no. 2, pp. research0009.1-0009.8, 2002.

[25] W. Pan, *Incorporating gene functions as priors in model-based clustering of microarray gene expression data*. Bioinformatics, vol. 22, no. 7, pp. 795-801, 2006.

[26] S. A. Ramsey, S. L. Klemm, D. E. Zak, K. A. Kennedy, V. Thorsson, B. Li, M. Gilchrist, E. S. Gold, C. D. Johnson, V. Litvak, G. Navarro, J. C. Roach, C. M. Rosenberger, A. G. Rust, N. Yudkovsky, A. Aderem, and I. Shmulevich, *Uncovering a Macrophage Transcriptional Program by Integrating Evidence from Motif Scanning and Expression Dynamics*, PLoS Computational Biology, vol. 4, no. 2, pp. e1000021, 2008.

[27] G. Schwarz, *Estimating the dimension of a model*. Annals of Statistics, vol. 6, pp. 461-464, 1978.

[28] E. Segal, H. Wang, and D. Koller *Discovering molecular pathways from protein interaction and gene expression data*, Bioinformatics, vol. 19, no. 1, pp. i264-i272, 2003.

[29] G. Sherlock *Analysis of large-scale gene expression data*, Briefings in Bioinformatics, vol. 2, no. 4, pp. 350-362, 2001.

[30] P. Smyth, *Model selection for probabilistic clustering using cross-validated likelihood*. Statistics and Computing, vol. 9, pp. 63-72, 2000.

[31] M. Symons, *Clustering criteria and multivariate normal mixtures*, Biometrics, vol. 37, pp. 35-43, 1981.

[32] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation*, Proceedings of the National Academy of Sciences of the United States of America, vol. 96, pp. 2907-2912, 1999.

[33] A. Taylor and D. J. Higham, *Contest: A controllable test matrix toolbox for MATLAB*. Genome Research, vol. 16, pp. 627-635, 2007.

[34] K. Tu, H. Yu, and Y. X. Li, *Combing gene expression profiles and protein-protein interaction data to infer gene functions*, International Journal of Biotechnology, vol. 124, no. 3, pp. 475-485, 2006.

[35] N. Tuncbag, T. Haliloglu, O. Keskin, *Correspondence between function and interaction in protein interaction network of Saccaromyces cerevisiae*. International Journal of Biomedical Sciences, vol. 1, no. 1, pp. 1306-1216, 2006.

[36] S. Vaithyanathan and B. Dom, *Model-based hierarchical clustering*, Proceedings of the 16th conference on Uncertainty in Artificial Intelligence, Stanford, California, USA, 30 June - July 3, 2000, pp. 599-608.

[37] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, *Model-based clustering and data transformation for gene expression data*, Bioinformatics, vol. 17, no. 10, pp. 977-987, 2001.

Data set 1		cluster 1				cluster 2				cluster 3			
gB	α	20	5	3	30	20	25	30	35	2	15	33	4
	β	2	15	33	4	20	25	30	35	20	5	3	30
bB	α	33	30	22	20	30	27	20	18	27	24	18	16
	β	30	33	20	22	27	30	18	20	24	27	16	18
gG	μ	5	-8	20	15	10	1	-20	0	-10	8	5	15
	σ	1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
bG_m	μ	3	15	5	11	2	13	6	9	1	14	7	10
	σ	1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
bG_v	μ	5	-8	20	15	10	1	-20	0	-10	8	5	15
	σ	10	20	30	25	10	20	30	25	10	20	30	25
Data set 2		cluster 1				cluster 2				cluster 3			
gB	α	20	5	3	30	20	25	30	35	2	15	33	4
	β	2	15	33	4	20	25	30	35	20	5	3	30
bB	α	33	30	22	20	30	27	20	18	27	24	18	16
	β	30	33	20	22	27	30	18	20	24	27	16	18
gG	μ	10		1		-20		0		-10	8	5	15
	σ	1		2		3		2.5		1	2	3	2.5
bG_m	μ	2		13		6		9		1	14	7	10
	σ	1		2		3		2.5		1	2	3	2.5
bG_v	μ	10		1		-20		0		-10	8	5	15
	σ	10		20		30		25		10	20	30	25
Data set 3		cluster 1				cluster 2				cluster 3			
gB	α	20		5		3		30		2	15	33	4
	β	2		15		33		4		20	5	3	30
bB	α	30		27		20		18		27	24	18	16
	β	27		30		18		20		24	27	16	18
gG	μ	5	-8	20	15	10	1	-20	0	-10	8	5	15
	σ	1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
bG_m	μ	3	15	5	11	2	13	6	9	1	14	7	10
	σ	1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
bG_v	μ	5	-8	20	15	10	1	-20	0	-10	8	5	15
	σ	10	20	30	25	10	20	30	25	10	20	30	25

Note: 'gB' and 'bB' each stands for 'beta' distributed data that are of 'good' and 'bad' quality respectively; 'gG', 'bG_m' and 'bG_v' each represents 'Gaussian' distributed data that are of 'good' quality and 'bad' quality with respect to close means and large variances respectively; '||' separate the parameters of different clusters, and '|' separate the parameters of different dimensions (2nd dimension) within the same cluster.

Table I
PARAMETERS OF BETA AND GAUSSIAN DISTRIBUTED DATA.

		All		F		C		P		
Model	Criterion	M1	M2	M1	M2	M1	M2	M1	M2	N
BMM	1∼4	0.2487	0.2945	0.3546	0.3484	0.3579	0.3479	0.3498	0.3423	4
GMM	1∼2	0.1889	0.1756	0.2640	0.3149	0.2924	0.3451	0.2740	0.3334	13
	3∼4	0.2112	0.1914	0.2955	0.3027	0.3239	0.3587	0.3019	0.3356	29
BGMM	1∼4	0.1351	0.1350	0.2369	0.2681	0.2847	0.3117	0.2506	0.2976	4
sBGMM$_{P1_a}$	1∼4	**0.0848**	**0.0710**	0.2128	**0.2021**	0.2503	**0.2483**	0.2290	**0.2307**	4
sBGMM$_{P1_b}$	1∼4	0.0913	0.0747	**0.1947**	0.2174	**0.2272**	0.2684	**0.2110**	0.2409	4
sBGMM$_{P2_a}$	1∼4	0.1740	0.1840	0.2911	0.3279	0.3173	0.3337	0.2963	0.3225	16
sBGMM$_{P2_b}$	1∼4	0.1506	0.1291	0.3000	0.3536	0.3218	0.3700	0.3098	0.3638	9
sBGMM$_{P2_c}$	1∼4	0.1817	0.1785	0.2429	0.2926	0.2697	0.3083	0.2556	0.3040	8
P2$_a$		0.1948	0.1810	0.2610	0.2530	0.2833	0.3035	0.2649	0.2609	8
P2$_b$		0.2055	0.2167	0.2707	0.2736	0.2970	0.3074	0.2768	0.3043	29
P2$_c$		0.2216	0.2286	0.2726	0.2577	0.2999	0.2938	0.2815	0.2862	31

Note: 'F', 'C', 'P' represent the three aspects of gene ontology, and 'All' means all three aspects are included. 'M1' and 'M2' stand for the mean and median of the p-values across all the clusters, respectively. 'Model' and 'Criterion' represent the model and model selection criteria, respectively. Subindexes of sBGMM indicate the prior that is used, e.g. sBGMM$_{P1_a}$ stands for using prior 'P1$_a$'. '1' to '4' each represents model selection criterion BIC, ICL, AIC, AIC3 respectively. 'N' means the number of clusters generated by each model. The last three lines show the corresponding statistics for the clusters given by DAVID. The smallest p-value in each column is shown in bold face. All fractions are rounded to four decimal points.

Table IV
PERFORMANCE TEST RESULTS OF SBGMM WITH REAL DATA.

		All		F		C		P		
Moldel	Crit	M1	M2	M1	M2	M1	M2	M1	M2	N
sBGMM$_{normal}$	1∼4	0.1662	0.1002	0.2356	0.2776	0.2797	0.3222	0.2441	0.2976	13
sBGMM$_{core}$	1∼4	**0.0557**	**0.0308**	**0.1418**	**0.1492**	**0.1922**	**0.1595**	**0.1617**	**0.1551**	5
BGMM$_{core}$	1∼4	0.1279	0.0714	0.2259	0.2701	0.2682	0.2833	0.2373	0.2637	8

Note: 'F', 'C', 'P' represent the three aspects of gene ontology, and 'All' means all three aspects are included. 'M1' and 'M2' stand for the mean and median of the p-values across all the clusters, respectively. 'Model' and 'Crit' represent the model and model selection criteria, respectively. 'bef' and 'aft' in the subindexes of sBGMM represent that the clustering is done before and after knowing the core TF genes, respectively, and the last digit 'i' ($i \in 1, \ldots, 3$)in the subindex represents the 'i^{th}' repetition of clustering with this model. '1' to '4' each represents model selection criterion BIC, ICL, AIC, AIC3 respectively. 'N' means the number of clusters generated by each model. The smallest p-value in each column is shown in bold face. All fractions are rounded to four decimal points.

Table V
CLUSTERING RESULTS WITH WHOLE DATA SET.