sBGMM: a stratified Beta-Gaussian 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
Email: harri.lahdesmaki@tut.fi

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

***Abstract*—Cluster analysis is widely applied to discover the function of previously unannotated genes. This paper presents a novel stratified Beta-Gaussian mixture model, sBGMM, for clustering genes based on gene expression data, protein-DNA binding data and data that can provide information for constructing priors such as protein-protein interaction (PPI) data. An expectation maximization (EM) type of algorithm for Beta mixture model is first developed and then combined with that of Gaussian mixture model. This combined algorithm can jointly estimate the parameters for both Beta and Gaussian distributions and is used as the core in the sBGMM method. The stratification property of sBGMM is exhibited as Stratum-specific prior probabilities and is constructed by the pre-cluster results obtained from PPI data in this study. This proposed sBGMM method differs from other mixture model based methods in its integration of two different data types into a single and unified probabilistic modeling framework and incorporation of prior information from a third data source. Several well-studied model selection methods, such as Akaike information criterion (AIC), modified AIC (AIC3), Bayesian information criterion (BIC), and integrated classification likelihood-BIC (ICL-BIC) are applied to estimate the number of clusters, and simulation results show that AIC3 works best for sBGMM. Simulations also indicate that combining two different data sources into a single mixture model can greatly improve the clustering accuracy and stability, and employing priors to stratify the model can further enhance its performance. This proposed method provides a more efficient use of multiple data sources than methods that analyze different data sources separately.**

I. Introduction

The attitude that different biological data types offer information from different perspectives and should be combined to make predictions more robust has become well accepted in the field of system biology. Thus how to integrate different data types to make the results more accurate has become one of the most challenging problems. In the case of gene clustering, among the studies done by many researchers, we have successfully incorporated protein-DNA binding data with expression data [6], and the significant increase in its prediction accuracy and stability compared with that obtained from single data sources demonstrated the power of integrating more information in the better understanding of the biological system and encouraged us to dig farther in this direction.

The previous work that we have done is to combine beta mixture model (BMM) and Gaussian mixture model (GMM) into a single model based framework, which is named Beta-Gaussian mixture model (BGMM), for gene clustering [6]. The results are satisfactory, however, there is still much room left for performance improvement. Inspired by the work done by Pan [9], where stratum-specific prior probabilities are put forward and verified to yield better results than uniform priors, we construct the stratified priors in this study by converting protein-protein interaction (PPI) data into contact matrix through which the genes (corresponds to proteins that it encodes in PPI data) are pre-clustered [13]. Simulations were done to compare the performance of sBGMM and its non-stratified form (BGMM), and the results support sBGMM because of its consistently higher clustering accuracy in all the tested cases.

Criteria for model selection can be classified into two main groups, which are likelihood-based methods and approximation-based methods [11]. Likelihood-based methods include cross-validation and bootstrap methods, and cross-validation method can further be divided into many different strategies with respect to how the partitions are chosen. As these methods are computationally more expensive, approximation-based methods are widely preferred by most people. These methods include penalized likelihood, closed-form approximations to the Bayesian solution, and Monte Carlo sampling of the Bayesian solution, among which penalized likelihood method is most prevalent. Penalized likelihood criteria mainly refers to Akaike information criterion (AIC), modified AIC (AIC3), Bayesian information criterion (BIC), integrated classification likelihood-BIC (ICL-BIC), and minimum description length (MDL). They are typically derived from approximations based on asymptotic arguments as the data size N approaches ∞ [11]. Thus approximation-based methods can suffer from theoretical limitations on their applicability to mixture problems in small-sample setting, and can dependent on the accuracy of the underlying approximations or simulations in a non-transparent manner [11]. AIC and BIC are often reported to work better than the other criteria for Gaussian mixture model (GMM) [3], [8]. ICL-BIC is preferred by Beta mixture model (BMM) according to [8]. Our previous work suggests using AIC for BMM and AIC3 for BGMM [6] through comparing four well-known approximation-based criteria (AIC, AIC3, BIC, ICL-BIC), and in this study we choose

AIC3 as the model selection criterion with the same strategy.

The following sections are organized as 'Methods', 'Results', and 'Conclusions'. Section 'Methods' is divided into two parts. In part one, mixture model based clustering and EM algorithm are discussed (including the short description of classic GMM, our BMM, and the joint model BGMM which belongs to our previous work). Part two of this section introduces how the criteria for model selection (AIC, AIC3, BIC, ICL-BIC) are formulated, and how the criterion for sBGMM is chosen. Section 'Results' also consists of two parts. We first compare the performance of sBGMM with its non-stratified form in dealing with bad quality data, and then evaluate and compare the abilities of sBGMM and BGMM in handling with Region 2 data (Regions are divided in Section III). In section 'Conclusions', we first summarize this study and discuss its possible extension and applications to other problems; then mention the limitations within the proposed method; and finally suggest the possible future work related to the proposed sBGMM.

II. Methods

A. Mixture model based clustering and EM algorithm

In model-based clustering methods, each observation x is drawn from a finite mixture distributions with the prior probability π_i, component-specific distribution $f_i^{(g)}$ and its parameters θ_i. The formula is given as

$$f(x;\Theta) = \sum_{i=1}^{g} \pi_i f_i^{(g)}(x;\theta_i), \tag{1}$$

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

EM algorithm is then derived for the above model-based clustering. The data log-likelihood can be written as

$$\log L(\Theta) = \sum_{j=1}^{n} \log\Big(\Big[\sum_{i=1}^{g} \pi_i f_i(x_j;\theta_i)\Big]\Big), \tag{2}$$

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

In order to make the maximization of Equation 2 tractable, the problem is casted in the framework of incomplete data. Define z_{ji} as the indicator of whether x_j is from component i, i.e., $z_{ji} = 1$ if x_j is from component i and $z_{ji} = 0$ otherwise. Then the complete data log-likelihood becomes

$$\log L_c(\Theta) = \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ji} \log\left(\pi_i f_i(x_j;\theta_i)\right). \tag{3}$$

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

$$\begin{aligned} Q(\Theta;\Theta^{(m)}) &= E_{\Theta^{(m)}}(\log L_c|X) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ji}^{(m)} \log\left(\pi_i f_i(x_j;\theta_i)\right), \end{aligned} \tag{4}$$

where $\Theta^{(m)}$ represents the parameter estimates at iteration m). M step updates the parameter estimates to maximize Q. The algorithm is iterated until convergence. Note that z's in Equation 3 are replaced with τ's in Equation 4, and the relationship between these two parameters is

$$\tau_{ji} = E[z_{ji}|x_j, \hat{\theta}_1, ..., \hat{\theta}_g, \hat{\pi}_1, ..., \hat{\pi}_g]. \tag{5}$$

The set of parameter estimates $\left\{\hat{\theta}_1, ..., \hat{\theta}_g, \hat{\pi}_1, ..., \hat{\pi}_g\right\}$ is a maximizer of the expected log-likelihood for given τ_{ji}'s, and we can assign each x_j to its component based on $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$. In the following subsections, GMM, BMM, BGMM, which are the component models of sBGMM, are first briefly described together with their corresponding EM algorithms (whose details can be found in our another paper [6]), and then sBGMM and its EM algorithm are proposed in detail.

1) GMM and its EM algorithm: The most widely used and well known model-based clustering method is finite GMM, in which each component is assumed to follow a Gaussian distribution. In this study we use the standard p dimensional normal distribution with mean μ_i and unconstrained covariance matrix V_i for each component in GMM. We run the EM algorithm multiple times with different initial values, where fuzzy c-means clustering algorithm is used for initialization, to avoid possible local maxima.

2) BMM and its EM algorithm: BMM is developed to tackle with data within boundaries $[0, 1]$ which is used for clustering protein-DNA binding data in this study. In this model, each component is assumed to be the product of p independent beta distributions, whose probability density function is defined as

$$f_i(x;\alpha_i,\beta_i) = \prod_{j=1}^{p} \frac{x^{\alpha_{ij}-1}(1-x)^{\beta_{ij}-1}}{B(\alpha_{ij},\beta_{ij})}. \tag{6}$$

In the EM algorithm, τ_{ji}'s are first calculated with current parameters, according to which x_j's are clustered to their corresponding clusters using z_{ji_0}'s (where $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$) in E-step. Then M-step is run to maximize Equation 3 instead of 4 (compared with the standard EM). In this step, new parameters $\hat{\alpha}$'s and $\hat{\beta}$'s are numerically estimated using the maximum likelihood principle (matlab function 'betafit' is used here for this purpose) given the hard clusters obtained in E-step, and the new $\hat{\pi}$'s are calculated by

$$\hat{\pi}_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ji}^{(m)}/n, \tag{7}$$

where

$$\tau_{ji}^{(m)} = \frac{\pi_i^{(m)} f_i(x_j; \alpha_i^{(m)}, \beta_i^{(m)})}{\sum_{i=1}^{g} \pi_i^{(m)} f_i(x_j; \alpha_i^{(m)}, \beta_i^{(m)})}. \quad (8)$$

The algorithm is run multiple times with randomized initial values.

3) BGMM and its EM algorithm: EM for BMM and GMM are combined into a single framework in BGMM with the assumption that, for each component i, the expression and binding data are independent. The procedures of parameter maximization for both data types are the same as those for BMM and GMM, except that the calculation of τ's is the product of two distributions

$$\tau_{ji}^{(m)} = \frac{\pi_i^{(m)} f_i^G(x_j; \mu_i^{(m)}, V_i^{(m)}) f_i^B(x_j; \alpha_i^{(m)}, \beta_i^{(m)})}{\sum_{i=1}^{g} \pi_i^{(m)} f_i^G(x_j; \mu_i^{(m)}, V_i^{(m)}) f_i^B(x_j; \alpha_i^{(m)}, \beta_i^{(m)})}. \quad (9)$$

Note that the superscripts (G) and (B) of f's mean that the parameters they represented are from GMM and BMM respectively.

4) sBGMM and its EM algorithm: In order to integrate as many data sources as possible, we proposed a sBGMM

$$f_{(k)}(x_j; \Theta_{(k)}) = \sum_{i=1}^{g} \pi_{(k),i} f_i^{(g)}(x_j; \theta_i), \quad (10)$$

where $1 \leq k \leq K$, which means that the genes can be partitioned into several groups, say $G_1, ...G_K$, according to certain criteria before EM is run.

In Equation 10, K stratified models share the same set of component distributions while differ in their usage of stratum-specific prior probabilities. Here in this study, this prior information is provided by the pre-cluster results obtained from PPI data [13]. PPI data is first converted into contact matrix (denoted as A) and then transformed into correlation matrix (denoted as C). Contact matrix is in the form of

$$A = \begin{cases} 1 & if \quad i \Leftrightarrow j \\ 0 & if \quad i \nLeftrightarrow j, \end{cases} \quad (11)$$

where $i \Leftrightarrow j$ means the existence of a connection between node i and j while $i \nLeftrightarrow j$ denotes the other way around. Before obtaining the correlation matrix, the pathlength between nodes i and j which is denoted as P_{ij} and characterised as the smallest integer $k \geq 1$ such that $(A^k)_{ij} \neq 0$ is calculated for all pairs of nodes. Here we imployed and modified the 'pathlength' function in the 'CONTEST' toolbox in matlab [12] to calculate P. If P_{max} was denoted as $\max_{ij} P_{ij}$ for $(A^k)_{ij} \neq 0$, then the modification is done to set $P_{ij} = P_{max}$ when $(A^k)_{ij} = 0$ for all k. The pathlength matrix P is then used to obtain the correlation matrix [2]

$$C(i,j) = 1 - \frac{P_{ij}}{P_{max}}. \quad (12)$$

We use the correlation 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. The 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. The priors given to each clusters in this study are randomly generated.

In E step, updates of μ's, V's, α's, and β's are the same as that in BGMM, but for πs they are updated by

$$\hat{\pi}_{(k),i}^{(m+1)} = \sum_{j \in G_k} \tau_{ij}^{(m)} / n_k, \quad (13)$$

where n_k is the number of the genes in G_k and m stands for the number of iterations.

In this study, for each data set we run each EM algorithm 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 convergences according to the statistics stored during the simulations.

B. Model Selection

Four well-known approximation-based model selection criteria, AIC [3], [1], AIC3 [3], [5], BIC [9], [10], and ICL-BIC [8], are compared in sBGMM, whose formulations are defined as

$$\begin{aligned} AIC &= -2\log L(\hat{\Theta}) + 2d, & (14)\\ AIC3 &= -2\log L(\hat{\Theta}) + 3d, & (15)\\ BIC &= -2\log L(\hat{\Theta}) + d\log(nM), & (16)\\ ICL-BIC &= -2\log L(\hat{\Theta}) + d\log(nM) \\ &\quad -2\sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ji}\log(\tau_{ji}), & (17) \end{aligned}$$

where d is the number of free parameters, and M in equations 16 and 17 is the total dimension of the data ($M = \sum_{w=1}^{W} M_w$, M_w is the dimension 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 $C_{ji} = (\tau_{ji})$ [8]. The number of free parameters d is different in different models. As described in [6], we have $d_G = (p^2+p)g/2 + pg + g - 1$, $d_B = 2gp + g - 1$, $d_{BG} = d_B + d_G - (g-1)$, where the subindices indicate their corresponding models. In sBGMM, there are K times π's as that in its non-stratified form, so we have $d_{sBG} = d_{sB} + d_{sG} - K(g-1)$.

III. Results

We compared the performance of sBGMM and BGMM with three artificial datasets, which are generated by a simplified model (we generate data from a diagonal covariance model although our model assumes unconstraint covariance). Data set 1 and 2 are the same as what was used in [6], where the first data set has the same parameters for all the dimensions within each cluster ('non-mixed' case) while the second data set does not ('mixed' case). The third data set is added in this study,

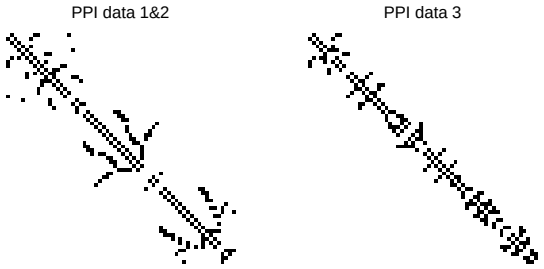

Figure 1. Sparsity patterns of the contact matrix of the designed PPI data. PPI 1&2 is used in Simulation 1 and 2, and PPI 3 is used for Simulation 3

which has different number of components for different data types (six components for binding data and three components for expression data). Each artificial data set was designed to fall into four categories: 'good Beta' (gB), 'bad Beta' (bB), 'good Gaussian' (gG), 'bad Gaussian' (bG), where 'good' stands for 'less noisy data', and 'bad' means the opposite. For the first two simulations, we also designed two kinds of 'bG', 'bG_m' and 'bG_v', which are hard to be clustered compared to 'gG' with respect to means and variances respectively. Parameter settings for the datasets are listed in Table I, where the combination of 'good Gaussian variance' and 'bad Gaussian mean' is 'bG_m', and the combination of 'good Gaussian mean' and 'bad Gaussian variance' is the case 'bG_v'. Two PPI datasets were designed to match the dimensionalities and underlying components of expression and binding data (PPI data for the third data set is designed to have six components), whose sparsity patterns are shown in Fig. 1. The dimensions of the data are $n = 60$ and $p = 4$ for all data sets, and all the simulations are repeated 20 times with randomly generated data sets.

We used the same scoring system as developed in [6] for evaluating the clustering accuracy, which is denoted as 'E score'

$$\begin{aligned} e_j(r) &= \begin{cases} 1 & if \quad \hat{z}_{ji} = 1 \quad and \quad r_i = T_j \\ 0 & otherwise \end{cases} \\ E &= \max_{r \in R} \sum_{j=1}^{n} e_j(r)/n \qquad (18) \\ R &= \big\{ r = (r_1, \ldots, r_{\hat{g}}) : \forall i \neq j \; r_i \neq r_j; \\ &\quad r_i \in \{1, \ldots, \max\{\hat{g}, g\}\} \big\}. \end{aligned}$$

In this scoring system, T_j denotes the ground truth clustering membership of data j, and r_i is the label of data belonging to component i predicted by the clustering algorithm; r is chosen from labels $1, 2, \ldots, \max\{\hat{g}, g\}$, where $\hat{g}$ and g are the largest labels in the estimated and ground truth clustering. Also note that e is the individual score of each gene, E is the average score of all the genes for each repetition, 'E score' of each repetition is the one corresponding to the optimal Q, and the final 'E score' of each data set is 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.

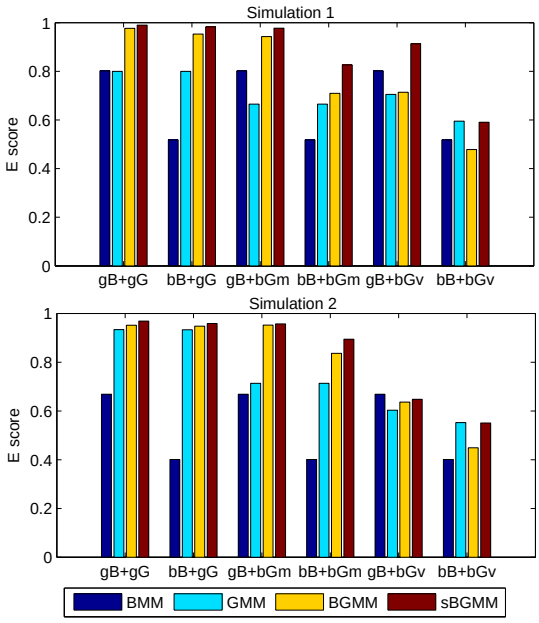

Figure 2. Results of Simulation 1 and Simulation 2. Performance comparison of sBGMM with all of its component models.

Comparison results of different model selection criteria in sBGMM are shown in Table II, according to which AIC3 is chosen for this model. AIC, BIC and AIC3 which are selected for BMM, GMM and BGMM respectively in [6] are also used in the following simulations.

Results of the first two simulations are shown in Fig. 2. In both simulations, E scores of sBGMM are consistently higher than those of BGMM. Although for Gaussian distributed data that have large variances the same problem for BGMM still lies in sBGMM (the performance is even lowered down in the joint models), it outweighs BGMM under all the tested circumstances. The performance improvement brought out by sBGMM compared with BGMM indicates that the clustering accuracy of BGMM can be increased by adding stratum-specific priors from another data source.

According to work done in [7], only partial gene expression and protein-DNA binding data agree with each other (consisting of the same number of clusters), and the data fall into three regions which are shown in Fig. 3. Data were designed to have the same number of clusters (Region 1) in the first two simulations, and adding stratified priors (sBGMM) to the joint model (BGMM) can consistently improve the clustering accuracy according to the simulated results obtained in this study. For data within Region 3, as expression data contains more information than binding data (expression data gives more clusters than binding data) and can already give good performance, there is no need to incorporate protein-DNA binding data in this case (we only use low quality Gaussian distributed data in this simulation). To test the performance of sBGMM and BGMM in handling with data within Region

Data		Data set 1			Data set 2												Data set 3					
		c1	c2	c3	c1				c2				c3				c1	c2	c3	c4	c5	c6
gB	α	10	20	25	15	20	25	20	20	25	15	5	1	20	1	30	10	20	1	20	1	100
	β	20	10	20	20	15	20	25	20	25	15	5	20	1	30	1	20	10	20	1	100	1
bB	α	10	15	17	15	10	25	20	10	5	15	12	30	25	30	35	10	20	15	20	17	18
	β	20	20	18	10	15	20	25	5	10	12	15	25	30	35	30	20	10	20	15	18	17
gG	μ	7	8	9	9	-9	11	-11	10	-10	12	-12	11	-11	13	-13	5		10		20	
	σ	0.3	0.4	0.2	0.7	0.2	0.7	0.2	0.8	0.3	0.8	0.3	0.9	0.4	0.9	0.4	0.5		0.8		0.1	
bG	μ	7.5	8	8.5	9.5	-9.5	10	-10	9	-9	9.5	-9.5	10	-10	9	-9	7		8		9	
	σ	1	0.9	0.8	1	1	1.5	1.5	1.5	1.5	2	2	2	2	1	1	0.3		0.4		0.2	

Note: 'gB' and 'bB' each stands for 'Beta' distributed data that are of 'good' and 'bad' quality respectively; 'gG' and 'bG' each represents 'Gaussian' distributed data that are of 'good' and 'bad' quality respectively; 'α', 'β', 'μ', 'σ' are the parameters in each corresponding distribution.

Table I
DATA SETS DESIGNED FOR SIMULATIONS

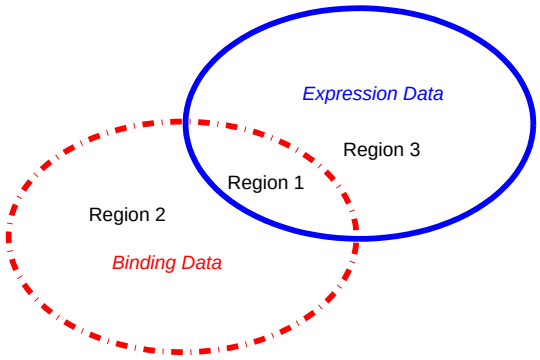

Figure 3. Region divisions of input data. In Region 1 gene expression and protein-DNA binding data have the same number of underlying components; in Region 2 binding data has more components; in Region 3 expression data has more components

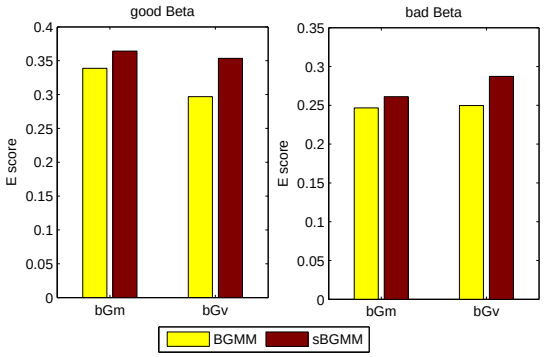

Figure 4. Results of Simulation 3. Performance test and comparison of 'sBGMM' and 'BGMM' in dealing with Region 2 data.

2, we did Simulation 3 where binding data and PPI data (both have six clusters) have more components than that of expression data (three clusters). According to the results (shown in Fig. 4), where the 'E score's are obtained with the assumption that the real number of underlying clusters was six, sBGMM outweighs BGMM. This result combined with those obtained from Simulation 1 and 2 has demonstrated the power of employing stratified priors for the joint model in improving the clustering accuracy.

IV. CONCLUSIONS

This paper presents a novel method based on stratified Beta-Gaussian mixture model, sBGMM, for gene clustering from multiple data sources. In this study, we integrated gene expression data, protein-DNA binding data and PPI data, where expression data and protein-DNA binding data are assumed to be of Gaussian and Beta distribution respectively, and PPI data is used to set the prior weights of genes belonging to each pre-cluster. An EM type of algorithm for estimating parameters from beta distribution is developed and combined with the classical EM for Gaussian distribution into a single framework. This joint EM algorithm is used as the core for sBGMM as well as its non-stratified form. sBGMM differs from BGMM in its stratified priors, where prior weights are set equal to genes belonging to the same cluster according to some prior information. In principle, this proposed sBGMM is not limited to the data we have used in this study. Any data that follows the assumption of Gaussian and Beta distribution could be integrated into this framework, and any information that can pre-cluster the data might serve to set the prior weights. This work demonstrates one approach of integrating information from multiple data sources. Data of other distributions (other than Gaussian and beta) can also be incorporated by joining EM algorithm of that particular distribution into this framework in a similar way. So in a sense, the method proposed in this paper is applicable to many problems and not limited to the particular problem considered here.

However, there are some limitations of this current work. First, for Gaussian distributed data that has large variances, the performance improvement of the joint model compared with BMM and GMM still can not be guaranteed. This might due to the less optimal pre-clustering method that we have used for constructing priors from PPI data, which renders the priors less informative for further clustering. Second, as we are using EM algorithm with arbitrary covariance matrices as the core for GMM based on which all the joint models are developed,

			sBGMM			
Data set	d		AIC	AIC3	BIC	ICL
1	2	gB+gG	0	0	0	0
		bB+gG	0	0	11	10
		gB+bG$_m$	0	0	1	1
		bB+bG$_m$	0	0	20	20
		gB+bG$_v$	0	0	18	18
		bB+bG$_v$	0	2	20	20
	3	gB+gG	0	16	20	20
		bB+gG	7	18	9	10
		gB+bG$_m$	0	15	19	19
		bB+bG$_m$	3	13	0	0
		gB+bG$_v$	0	13	2	2
		bB+bG$_v$	1	11	0	0
	4	gB+gG	11	4	0	0
		bB+gG	7	2	0	0
		gB+bG$_m$	10	5	0	0
		bB+bG$_m$	8	7	0	0
		gB+bG$_v$	9	7	0	0
		bB+bG$_v$	11	7	0	0
	5	gB+gG	9	0	0	0
		bB+gG	6	0	0	0
		gB+bG$_m$	10	0	0	0
		bB+bG$_m$	9	0	11	11
		gB+bG$_v$	11	0	0	0
		bB+bG$_v$	8	0	0	0
2	2	gB+gG	0	0	0	0
		bB+gG	0	1	7	7
		gB+bG$_m$	0	0	4	4
		bB+bG$_m$	0	1	18	18
		gB+bG$_v$	0	0	20	20
		bB+bG$_v$	2	9	20	20
	3	gB+gG	0	12	20	20
		bB+gG	15	17	13	13
		gB+bG$_m$	0	11	16	16
		bB+bG$_m$	12	16	2	2
		gB+bG$_v$	0	9	0	0
		bB+bG$_v$	5	5	0	0
	4	gB+gG	9	8	0	0
		bB+gG	4	2	0	0
		gB+bG$_m$	6	7	0	0
		bB+bG$_m$	8	3	0	0
		gB+bG$_v$	5	8	0	0
		bB+bG$_v$	11	6	0	0
	5	gB+gG	11	0	0	0
		bB+gG	1	0	0	0
		gB+bG$_m$	14	2	0	0
		bB+bG$_m$	0	0	0	0
		gB+bG$_v$	15	3	0	0
		bB+bG$_v$	2	0	0	0

Note: Values shown here is the number of occurrence out of the total tests. Notation 'd' means the number of components; 'ICL' is short for 'ICL-BIC'.

Table II
COMPARISON OF DIFFERENT MODEL SELECTION CRITERIA IN SBGMM

the inherent drawbacks of this algorithm also exist in our joint methods. For large-dimensional data, the complexity of the model increases dramatically and might result in selecting less optimal model and slow convergence [4].

For future work, we have two directions. First, we could integrate more data types into the proposed mixture model framework. In this direction, the most obvious start might be to incorporate PPI data as part of the joint statistical mixture model instead of using it only to construct the prior, and the prior might be obtained from Gene Ontology (GO) information. Second, we will apply our method to real data. In this direction, many possible work could be done, including developing more robust method for pre-clustering PPI data, reducing the dimensionality of the data, implementing better EM algorithm into our framework (e.g. implement EM with diagonal covariance matrix for GMM; develop the standard EM for BMM), and applying likelihood-based criteria for model selection in sBGMM such as cross-validated likelihood method [11].

ACKNOWLEDGMENT

We would like to thank the Tampere Graduate School in Information Science and Engineering (TISE) for its financial support in this project.

REFERENCES

[1] H. Akaike, *A new look at the statistical identification model*. IEEE Transactions on Automatic Control, vol. 19, pp. 716-723, 1974.

[2] S. Asur, D. Ucar, S. Parthasarathy, *An ensemble framework for clustering protein-protein interaction networks*. Bioinformatics, vol. 23, pp. i29-i40, 2007.

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

[4] N. Bouguila and D. Ziou, *A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized Dirichlet mixture*. IEEE Transactions on Image Processing, vol. 15, no. 9, pp. 2657-2668, 2006.

[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, H. Lähdesmaki, 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, 2008, accepted.

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

[8] 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.

[9] 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.

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

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

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

[13] 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.