# BGMM: A BETA-GAUSSIAN MIXTURE MODEL FOR CLUSTERING GENES WITH MULTIPLE DATA SOURCES

*Xiaofeng Dai, Harri Lädesmäki, Olli Yli-Harja*

Department of Signal Processing,
Tampere University of Technology, Tampere, Finland
{xiaofeng.dai,harri.lahdesmaki,olli.yli-harja}@tut.fi

## ABSTRACT

This paper presents a novel Beta-Gaussian mixture model, BGMM, for clustering genes based on gene expression data and protein-DNA binding 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 BGMM method. Four well-studied model selection methods, 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 AIC3 works best for BGMM in our simulations. Simulations also indicate that combining two different data sources into a single mixture model can greatly improve the clustering accuracy and stability. The proposed BGMM method differs from other mixture model based methods in its integration of two different data types into a single and unified probabilistic modeling framework, which provides a more efficient use of multiple data sources than methods that analyze different data sources separately.

## 1. INTRODUCTION

It has become more and more acknowledged that different data sources offer information from different aspects, and their combination can make the prediction more robust. Thus how to integrate different data types to make the results more accurate has become one of the most challenging problems in the field of system biology. In the context of gene clustering, gene expression data has been widely used with the assumption that genes which have similar expression pattern under different conditions have similar cellular functions, are likely to be involved in the same cellular processes [5]. This assumption might be too ideal considering the complexity of real biological systems. However, if we could incorporate physical binding information, such as the probabilities of certain binding events occurring among gene products and genes (protein-DNA binding data), into expression data based clustering framework, the clustering results might be more trustable with respect to similar cellular functions, processes and co-regulation. In this study, we developed a clustering algorithm which can cluster genes based on their expression data and protein-DNA binding data.

Many unsupervised methods have been developed and widely used in gene clustering. They can be roughly classified into three categories, which are heuristic, iterative relocation and model-based methods [3]. The first two approaches have problems with solving some basic practical issues such as 'how to define the number of clusters' and 'how to handle outliers'. In model-based methods, the first question can be recasted as the model selection problem. For the second problem, the outliers can be handled by adding one or more components which represent a different distribution for them [3, 4]. Moreover, model-based clustering methods outweigh approaches within the other two categories in their statistical nature [3]. So in this study, we choose model-based clustering as the framework for unsupervised data fusion.

Expectation maximization (EM) algorithm is generally used to solve the problem of maximum likelihood estimation with incomplete data, and thus is commonly adopted in model-based clustering. Although EM algorithm for Gaussian distribution is well-known, less information is available about EM algorithm for other distributions, not mentioning combinations of different distributions. In our study, gene expression data and protein-DNA binding data are integrated into a combined mixture model. We first developed an EM type of algorithm for beta distribution, and then combined it with that for Gaussian distribution. Simulation results show that our joint mixture model can yield better results compared with either of its component models, which demonstrates the idea that the more data that are integrated the better the result turns out to be.

Criteria for model selection can be classified into likelihood-based methods and approximation-based methods, of which approximation-based methods are widely preferred by its simplicity and less computational cost [9]. 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. Four well-known penalized likelihood criteria, Akaike information criterion (AIC), modified AIC (AIC3), Bayesian information criterion (BIC), and integrated classification likelihood-BIC

(ICL-BIC) were tested in BGMM and its component models (Beta mixture model 'BMM', Gaussian mixture model 'GMM') in this study. AIC and BIC are commonly used as the criterion for GMM [1, 4], and ICL-BIC is reported to work better for BMM according to [4]. Our simulation results suggest using AIC and AIC3 in BMM and BGMM respectively and embrace the tradition of employing BIC in GMM.

The following sections are organized as 'Methods', 'Results', and 'Conclusions'. Section 'Methods' is divided into two parts. In the first part, mixture model based clustering and EM algorithm are discussed, where the classic EM for GMM, our EM for BMM, and the joint EM for BGMM are all introduced. The second part of this section introduces the formulation of four tested model selection criteria (AIC, AIC3, BIC, ICL-BIC), and how the optimal criteria for each model was chosen. In section 'Results', we evaluated and compared the performance of BGMM with BMM and GMM. In section 'Conclusions', we summarized this study and discuss its possible extension and applications to other problems, and mentioned the possible future work that is related to the proposed BGMM.

## 2. METHODS

### 2.1. Mixture model based clustering and EM algorithm

In model-based clustering method, each observation $x$ is drawn from a finite mixture distributions with the prior probability $\pi_i$, component-specific distribution $f_i$ and its parameters $\theta_i$. The formula is given as

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

where $\Theta = \{(\pi_i, \theta_i) : i = 1, \dots, g\}$ is used to denote all unknown parameters, with the restriction that $0 \leq \pi_i \leq 1$ for any $i$ and that $\sum_{i=1}^{g} \pi_i = 1$. Note that $g$ is the number of components in this model.

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

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

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 indeed 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 which is denoted as $Q$

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

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 $\tau$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 $\tau_{ji}$s, and we can assign each $x_j$ to its component based on $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$.

#### *2.1.1. 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 $\mu_i$ and unconstrained covariance matrix $V_i$ for each component in GMM [6]. 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.

#### *2.1.2. BMM and its EM algorithm*

In order to make the model-based method to work for data within boundaries $[0, 1]$, we developed a BMM with the assumption that each component is a product of independent beta distributions. The probability density function is defined as

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

The details of our EM type of algorithm for BMM is described below. First, initialize the parameters. $\alpha$s and $\beta$s for each component beta distribution $k$ ($k \in \{1, \dots, p\}$) are initialized by method-of-moments so that their means are randomly distributed within the range of $x_{1k}, \dots, x_{nk}$ and variances are equal for all clusters ($g$); and for $\pi_i$s, they are initialized with the uniform probability $1/g$. Second, run E-step. Calculate $\tau_{ji}$ with current parameters, according to which $x_j$s are clustered to their corresponding clusters using $z_{ji_0}$s (where $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$). Third, run M-step to maximize Equation 3. Given the hard clusters obtained in E-step, numerically estimate the new parameters $\hat{\alpha}$s and $\hat{\beta}$s using the maximum likelihood principle (matlab function 'betafit' is used here for this purpose), and calculate the new $\hat{\pi}$s by

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

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

### 2.1.3. *BGMM and its EM algorithm*

EMs 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 $\tau$s is the product of two distributions

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

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

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.

## 2.2. Model Selection

Four well-known approximation-based model selection criteria, AIC [1], AIC3 [1, 2], BIC [7, 8], and ICL-BIC [4] are compared in BGMM and its component models, according to which the optimal criterion for each model is chosen. Calculations for the above criteria are defined in

$$\begin{aligned} AIC &= -2\log L(\hat{\Theta}) + 2d, && (10)\\ AIC3 &= -2\log L(\hat{\Theta}) + 3d, && (11)\\ BIC &= -2\log L(\hat{\Theta}) + d\log(nM), && (12)\\ ICL-BIC &= -2\log L(\hat{\Theta}) + d\log(nM) && \\ & \quad -2\sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ji}\log(\tau_{ji}), && (13) \end{aligned}$$

where $d$ is the number of free parameters in its corresponding model, and $M$ in equations 12 and 13 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 $C_{ji} = (\tau_{ji})$ [4].

The number of free parameters $d$ are different in different models. In GMM, we have $(p^2+p)g/2$ $\sigma$s, $pg$ $\mu$s, and $g-1$ free $\pi$s ($\sum_{i=1}^{g}\pi_i = 1$), so $d_G = (p^2+p)g/2 + pg + g - 1$. In BMM, as we have $pg$ $\alpha$s, $pg$ $\beta$s, and also $g-1$ free $\pi$s, $d_B = 2gp + g - 1$. In the joint model, the number of free parameters is the sum of those in its parents' models minus one set of free $\pi$s, thus we have $d_{BG} = d_B + d_G - (g-1)$.

## 3. RESULTS

In this study, we compared the performance of BMM, GMM and BGMM using two artificial datasets, which are generated by a simplified model (we generate data from a diagonal covariance model although our model assumes unconstraint covariance). Both datasets are designed to have three clusters and 60 by 4 dimensions ($n = 60$, $p = 4$). Parameters for different dimensions within each cluster are the same in the first data set but different in the second one, called 'non-mixed' and 'mixed' cases respectively. We designed two kinds of data for each data type within each data set, namely 'gB', 'bB', 'gG' and 'bG', which are short for 'good Beta' (less noisy, Beta distribution), 'bad Beta' (more noisy, Beta distribution), 'good Gaussian' (less noisy, Gaussian distribution), and 'bad Gaussian' (more noisy, Gaussian distribution) respectively. 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 1, 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$_v$'. All the simulations are repeated 20 times with randomly generated data sets.

In order to choose the optimal model selection criterion (with the highest score) for each model, we summed up the number of hits of the correct number of clusters for each data combination in both simulations. The summation results for AIC, AIC3, BIC, and ICL are 93, 71, 16 and 10 respectively in BMM, 8, 54, 64, 58 respectively in GMM, and 35, 101, 43, 43 respectively in BGMM, according to which AIC, BIC and AIC3 are chosen as the criteria for BMM, GMM, and BGMM respectively.

We developed one scoring system for evaluating the clustering accuracy, which is denoted as 'E score'

$$\begin{aligned} e_j(r) &= \begin{cases} 1 & if\ \hat{z}_{ji} = 1\ and\ r_i = T_j \\ 0 & otherwise \end{cases} \\ E &= \max_{r \in R} \sum_{j=1}^{n} e_j(r)/n \\ R &= \{r = (r_1, \ldots, r_{\hat{g}}) : \forall i \neq j\ r_i \neq r_j; \\ & \qquad r_i \in \{1, \ldots, \max\{\hat{g}, g\}\}\}, \end{aligned} \tag{14}$$

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.

| Data | | | Data set 1 | | | Data set 2 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | c1 | c2 | c3 | c1 | | | | c2 | | | | c3 | | | |
| Beta | good | alpha | 10 | 20 | 25 | 15 | 20 | 25 | 20 | 20 | 25 | 15 | 5 | 1 | 20 | 1 | 30 |
| | | beta | 20 | 10 | 20 | 20 | 15 | 20 | 25 | 20 | 25 | 15 | 5 | 20 | 1 | 30 | 1 |
| | bad | alpha | 10 | 15 | 17 | 15 | 10 | 25 | 20 | 10 | 5 | 15 | 12 | 30 | 25 | 30 | 35 |
| | | beta | 20 | 20 | 18 | 10 | 15 | 20 | 25 | 5 | 10 | 12 | 15 | 25 | 30 | 35 | 30 |
| Gaussian | good | mean | 7 | 8 | 9 | 9 | -9 | 11 | -11 | 10 | -10 | 12 | -12 | 11 | -11 | 13 | -13 |
| | | variance | 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 |
| | bad | mean | 7.5 | 8 | 8.5 | 9.5 | -9.5 | 10 | -10 | 9 | -9 | 9.5 | -9.5 | 10 | -10 | 9 | -9 |
| | | variance | 1 | 0.9 | 0.8 | 1 | 1 | 1.5 | 1.5 | 1.5 | 1.5 | 2 | 2 | 2 | 2 | 1 | 1 |

Table 1. Data sets designed for simulations

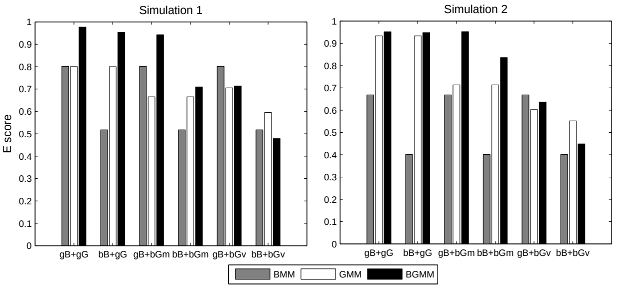

Figure 1. Performance test of BGMM.

The comparison results of BGMM with its component models are shown in Fig. 1. For expression data whose variances are not too large, the joint model can improve the clustering accuracy regardless of the quality of the data compared with either of its component models (E scores for cases 'gB+gG', 'bB+gG', 'gB+bG$_m$' and 'bB+bG$_m$' in BGMM are higher than those in GMM or BMM). However, if the expression data contains too much noise with respect to large variances('gB+bG$_v$', 'bB+bG$_v$'), the joint model does not necessarily yield better results. These results indicate that BGMM has the power of reinforcing each component model with information from the other one in both mixed and non-mixed cases but is sensitive to the variances of the Gaussian distributed data.

## 4. CONCLUSIONS

This paper presents a novel method based on Beta-Gaussian mixture model, BGMM, for gene clustering from multiple data sources. In this study, we integrated gene expression data and protein-DNA binding data, where expression data and protein-DNA binding data are assumed to be of Gaussian and Beta distribution respectively. An EM type of algorithm for estimating parameters from beta distribution is developed and combined with the EM for Gaussian distribution into a single framework, which is used as the core of BGMM. In principle, this proposed BGMM is not limited to the data we have used here, and any data that can be modeled as Gaussian and Beta distribution could be integrated into this framework. This work demonstrates one approach of integrating information from multiple data sources. Data of other distributions can also be incorporated by joining EM algorithm of that particular distribution into this framework in a similar way. Therefore BGMM is applicable to many problems and not limited to the particular problem considered here.

For future work, we will first apply our method to real data, where a possible problem might be the time issue due to the large dimensions of the data. Many techniques might be used to handle these problems such as reducing the dimension of the data or employing a faster EM framework. Second, we will integrate more data types into the proposed mixture model framework, where the most obvious start is to develope a stratified BGMM [7] which could incorporate one more data source by constructing the priors from a third data type.

## 5. REFERENCES

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

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

[3] 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.

[4] 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.

[5] D. X. Jiang, C. Tang, 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.

[6] G. Mclachlan and D. Peel, "Finite mixture models", *John Wiley & Sons*, 2000.

[7] 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.

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

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