IPSJ Transactions on Bioinformatics Vol. 2 47–62 (May 2009)

Original Paper

A Constrained Gaussian Mixture Model for Correlation-Based Cluster Analysis of Gene Expression Data

Naoto Yukinawa,[†1] Taku Yoshioka,[†2]
Kazuo Kobayashi,[†3] Naotake Ogasawara[†3]
and Shin Ishii[†1]

Clustering is a practical data analysis step in gene expression-based studies. Model-based clusterings, which are based on probabilistic generative models, have two advantages: the number of clusters can be determined based on statistical criteria, and the clusters are robust against the observation noises in data. Many existing approaches assume multi-variate Gaussian mixtures as generative models, which are analogous to the use of Euclidean or Mahalanobis type distance as the similarity measure. However, these types of similarity measures often fail to detect co-expressed gene groups. We propose a novel probabilistic model for cluster analyses based on the correlation between gene expression patterns. We also propose a "meta" cluster analysis method to eliminate the dependence of the clustering result on initial values of the clustering algorithm. In empirical studies with a time course gene expression dataset of *Bacillus subtilis* during sporulation, our method acquires more stable and informative results than the ordinary Gaussian mixture model-based clustering, k-means clustering and hierarchical clustering algorithms, which are widely used in this field. In addition, with the meta-cluster analysis, biologically-meaningful expression patterns are extracted from a set of clustering results. The constraints in our model worked more efficiently than those in the previous studies. In our experiment, such constraints contributed to the stability of the clustering results. Moreover, the clustering based on the Bayesian inference was found to be more stable than those by the conventional maximum likelihood estimation.

1. Introduction

Genome-wide gene expression profiling by microarrays provide quantitative transcriptional activation levels of thousands of genes at once. The measurements in multiple biological conditions or the time-course during various biological processes help us to reveal the corresponding genomic activities. Based on the primitive assumption that functionally related genes exhibit similar expression patterns, various cluster analysis (or simply clustering) methods have been developed and applied for microarray data.

Clustering algorithms can be roughly classified into two major categories: discriminative (or similarity-based) approaches and generative (or probabilistic model-based) approaches [1]. The former includes popular hierarchical clusterings [2),3)], self-organizing maps (SOMs) [4)–6)], k-means [7)], fuzzy C-means [8)], and Fuzzy ART [9)]. The latter type of method is referred to as "model-based clustering", which is based on a particular probabilistic generative model, i.e., a parametric family of probabilistic distributions, including multi-variate Gaussian mixture [10)–12)], mixture of t-distributions, mixtures of factor analyzers [13)], and von Mises-Fisher (vMF) mixtures [14)]. In addition to these approaches, stabilizing algorithms for arbitrary clustering methods have also been proposed [15)–17)].

In order to obtain biologically meaningful results from microarray datasets with a clustering method, it is desirable that the clustering method provides i) an appropriate similarity measure or the corresponding generative model for identifying co-regulated gene groups, ii) a determination method for the number of clusters by an objective criterion, and iii) robustness and stability against measurement noise that is inevitably contained in gene expression data. Most of the discriminative approaches above have the first property, since they can easily incorporate an arbitrary similarity measure. Meanwhile, the generative approaches often miss this property but have the second and the third properties.

A recent evaluation study of several popular gene clustering methods, including hierarchical clustering, k-means, k-medoids [18)], SOMs, multi-variate Gaussian mixture model-based clustering, and tight clustering [17)], showed that model-based clustering and tight clustering performed overall better than the other methods [19)]. This result suggests the importance of the second and third properties. However, similarity measures are still important because they not only have a direct effect on clustering outcomes but also define implicitly how the expressions of functionally-related genes behave.

In gene expression analyses, correlation-based similarity is more appropriate

†1 Graduate School of Informatics, Kyoto University
†2 ATR Computational Neuroscience Laboratories
†3 Graduate School of Information Science, Nara Institute of Science and Technology

than the Euclidean distance for extracting co-regulated genes because it is sensitive to relative variations of expression patterns rather than their magnitudes [14], but generative model-based clustering suitable to deal with such similarity has not been well studied.

In this study, we propose a cluster analysis procedure for gene expression data, based on the above background. For satisfying the first property, we introduce a probabilistic mixture model referred to as the constrained Gaussian mixture (CGM) model, which enables us to perform a correlation-based cluster analysis. A representative expression pattern within a cluster is incorporated as one of the model's parameters. The magnitude of each expression pattern is treated as a hidden variable. For the second property, we take an alternative approach to conventional maximum likelihood estimation or maximum *a posteriori* (MAP) estimation; parameters and hidden variables of the model are estimated by a variational Bayes (VB) method [20),21)], which approximately conducts a Bayesian inference and is effective in obtaining a statistically appropriate number of clusters. Furthermore, for removing the initial-condition problem and enhancing the stability of the clustering method, we propose a meta-cluster analysis algorithm that integrates the clustering results obtained in the previous (clustering) step.

The proposed method is applied to a time-course gene expression dataset of *Bacillus subtilis* during sporulation. In our experiment, the proposed probabilistic model and meta-cluster analysis produce clustering results that exhibit higher qualities than those by k-means and hierarchical clustering.

2. A Generative Approach for Correlation-Based Clustering

2.1 Constrained Gaussian Mixture Model

When gene expression levels are measured at D time points (or in D different conditions), the expression pattern of a single gene is represented by a D-dimensional vector. Suppose that the expression pattern $\boldsymbol{y}(i)$ of the i-th gene is generated by a noisy linear transformation:

$$\boldsymbol{y}(i) = \boldsymbol{w}x(i) + \boldsymbol{\epsilon}(i), \tag{1}$$

where $\boldsymbol{w}$ is a D-dimensional vector. $\boldsymbol{\epsilon}(i)$ is a random vector that obeys a D-dimensional normal distribution, $\mathcal{N}(\boldsymbol{\epsilon}(i)|\boldsymbol{0}, \sigma\boldsymbol{I}_D)$, whose mean and covariance are $\boldsymbol{0}$ and $\sigma^{-1}\boldsymbol{I}_D$, respectively. $\boldsymbol{I}_D$ denotes a D-by-D unit matrix. $x(i)$ is a random scalar that obeys a normal distribution, $\mathcal{N}(x(i)|\mu, 1)$, and is treated as a hidden variable. $\boldsymbol{w}$ and $x(i)$ are regarded as the representative expression pattern and the magnitude of the i-th gene expression pattern, respectively. When there are genes whose expression patterns are generated from Eq. (1), those patterns are correlated with the common representative expression pattern $\boldsymbol{w}$. It is found that $\boldsymbol{y}(i)$ obeys the following normal distribution:

$$P(\boldsymbol{y}(i)|\theta) = \mathcal{N}(\boldsymbol{y}(i)|\mu\boldsymbol{w}, (\sigma^{-1}\boldsymbol{I}_D + \boldsymbol{w}\boldsymbol{w}')^{-1}), \tag{2}$$

where $\theta \equiv \{\boldsymbol{w}, \sigma, \mu\}$ denotes the set of model parameters.

We consider a mixture model whose components are each defined by the probabilistic distribution (2) with different parameters. Let M be the number of components. The generative process of the mixture model begins by selecting a component. Then, an expression pattern is generated from the selected component. For convenience in explanation, we define an M-dimensional binary vector $\boldsymbol{z}(i) \equiv (z_1(i), \ldots, z_M(i))$ that satisfies $z_m(i) \in \{0, 1\}$ and $\sum_{m=1}^{M} z_m(i) = 1$. $z_m(i) = 1$ signifies that the m-th component is selected for the i-th gene. Since we do not know the correct answer to the clustering problem, $\boldsymbol{z}(i)$ is a hidden variable. We also use the notation $\boldsymbol{x}(i) \equiv (x_1(i), \ldots, x_M(i))$. Using these notations, the probabilistic distribution of a complete datum $(\boldsymbol{y}(i), \boldsymbol{x}(i), \boldsymbol{z}(i))$ is given by

$$P(\boldsymbol{y}(i), \boldsymbol{x}(i), \boldsymbol{z}(i)|\Theta) = \prod_{m=1}^{M} \{P(\boldsymbol{y}(i), x_m(i)|m, \theta_m) g_m\}^{z_m(i)}, \tag{3}$$

where $P(\boldsymbol{y}(i), x_m(i)|m, \theta_m)$ and θ_m denote the probabilistic distribution based on Eq. (2) and the set of parameters, respectively, of the m-th component. $\boldsymbol{g} \equiv (g_1, \ldots, g_M)$ denotes the mixing rate parameter that satisfies $g_m \geq 0$ ($m = 1, \ldots, M$) and $\sum_{m=1}^{M} g_m = 1$. $\Theta \equiv \{\{\theta_m\}_{m=1}^{M}, \boldsymbol{g}\}$ is the set of the mixture model's parameters.

The probabilistic distribution (2) is a normal (Gaussian) distribution that has constraints on its mean and covariance. Thus, the mixture model (Eq. (38)) is called a constrained Gaussian mixture (CGM) model. **Figure 1** shows the difference between a CGM model and a conventional Gaussian mixture model. If expression patterns are highly correlated with each other, those patterns are assigned to the same cluster regardless of their magnitude. On the other hand,

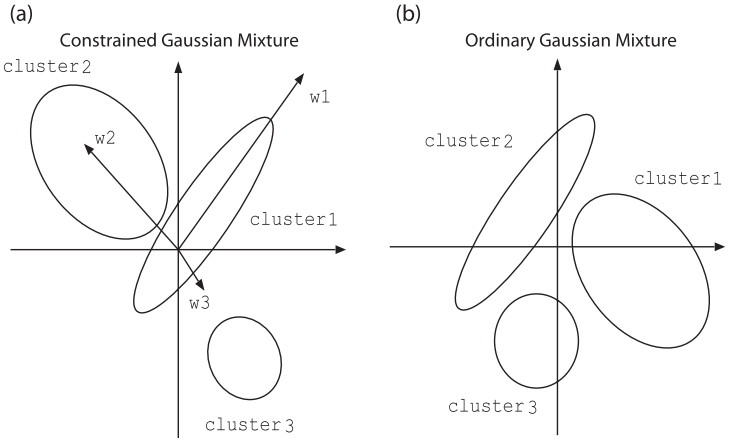


Fig. 1 CGM model and Gaussian mixture model. Conceptual illustration of (a) the proposed constrained Gaussian mixture (CGM) model and (b) the ordinary Gaussian mixture model. Ellipses represent covariances of clusters. The arrows in (a) represent the $\boldsymbol{w}$ vectors of the clusters, which are constrained to go through the origin.

a conventional Gaussian mixture model can divide such expression patterns into different clusters due to the large representation ability of the model (i.e., too many parameters).

The CGM model is a special case of the sparse coding methods [22)–24)], which is motivated from independent component analysis (ICA). In the formulation of ICA, an expression pattern $\boldsymbol{y}$ is represented by a linear mixture of unobservable sources $\boldsymbol{x}$: $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$, where $\boldsymbol{W}$ is the mixing matrix. The number of columns of $\boldsymbol{W}$ corresponds to the number of sources. Sparse coding means that only a few components of $\boldsymbol{x}$ take non-zero values. The CGM model can be regarded as the simplest case of "noisy" sparse coding: only one component in $\boldsymbol{x}$ takes a non-zero value, assuming the noise (see Eq. (1)). Thus, the CGM model is appropriate for clustering highly noisy data but is not adequate for problems like blind source separation, in which two or more sources are mixed.

The noise level often depends on the absolute expression level of genes [25),26)]. Such dependence should be incorporated into the probabilistic model when dealing with expression data consisting of absolute expression levels: $\boldsymbol{\epsilon}$ should depend on x in such a case. On the other hand, the dependence can be regarded as weak for data consisting of relative expression levels. The noise level for the relative expression value of a gene can be estimated in the preprocessing step, where the absolute expression values are used for calculating the relative expression value [27)]. Although we assume the normal distribution for noise $\boldsymbol{\epsilon}$ for simplicity, *a priori* knowledge about the noise level can be incorporated as a prior distribution within the Bayesian framework.

2.2 Parameter Estimation and Model Selection

Given expression patterns of N genes $Y \equiv \{\boldsymbol{y}(i)\}_{i=1}^{N}$, the likelihood for a CGM model with M clusters is calculated by

$$L(\Theta|Y;M) \equiv P(Y|\Theta,M) = \int P(Y,X,Z|\Theta,M)dXdZ, \tag{4}$$

where $P(Y,X,Z|\Theta)$ is the probability that the complete dataset $\{Y, X \equiv \{\boldsymbol{x}(i)\}_{i=1}^{N}, Z \equiv \{\boldsymbol{z}(i)\}_{i=1}^{N}\}$ is generated, which is given by Eq. (38). A maximum likelihood (ML) estimation obtains the parameter set that maximizes the likelihood Eq. (4). An ML estimation for models with hidden variables can be performed by the expectation-maximization (EM) algorithm [28)].

The likelihood can be used for obtaining the most probable parameters for a given model, while the Bayesian inference allows us to obtain the most probable model structure [29),30)]. Namely, we can determine an appropriate number of clusters within a Bayesian inference instead of the ML approaches which use some statistical criteria such as the Akaike information criterion (AIC) [31)] and the Bayesian information criterion (BIC) [32)]. The joint posterior distribution of unknown variables, model parameters and hidden variables, is considered in a Bayesian inference. According to the Bayes theorem, the joint posterior distribution of unknown variables in the CGM is given by

$$P(\Theta,X,Z|Y,M) = \frac{P(Y,X,Z|\Theta,M)P_0(\Theta|M)}{P(Y|M)} \tag{5}$$

$$P(Y|M) = \int P(Y,X,Z|\Theta,M)P_0(\Theta|M)d\Theta dXdZ, \tag{6}$$

where $P_0(\Theta|M)$ is the prior distribution of the parameters, which represents *a priori* knowledge of the parameters. The normalization factor $P(Y|M)$, called the marginal likelihood, represents the likelihood of the model (structure) with

M clusters. Thus, an ML estimation for the cluster number can be performed by the maximization of this marginal likelihood. For description simplicity, however, the cluster number M is omitted below.

The posterior distribution of Z is obtained by integrating out X and Θ from the joint posterior distribution $P(\Theta, X, Z|Y)$:

$$P(Z|Y) = \int P(\Theta, X, Z|Y) d\Theta dX. \tag{7}$$

If the marginalized posterior distribution (7) is obtained, the cluster index to which the i-th gene belongs is determined by $\mathrm{argmax}_m P(z_m(i) = 1|\boldsymbol{y}(i))$.

Since analytical calculation of the posterior distribution (5) is intractable for the CGM model, an approximate inference method, such as the Markov Chain Monte Carlo [33] (MCMC) or Laplace approximation [34], is needed. In this study, we use the variational Bayes (VB) method [20),21)].

We consider a trial posterior distribution $Q(\Theta, X, Z)$ that approximates the true posterior distribution $P(\Theta, X, Z|Y)$, and the free energy is defined by

$$\begin{aligned}\mathcal{F}[Q(\Theta, X, Z)] &\equiv \int Q(\Theta, X, Z) \ln \frac{P(Y, X, Z|\Theta) P_0(\Theta)}{Q(\Theta, X, Z)} d\Theta dX dZ \\ &= \ln P(Y) - \mathrm{KL}\left\{Q(\Theta, X, Z) \parallel P(\Theta, X, Z|Y)\right\}, \end{aligned} \tag{8}$$

where $\mathrm{KL}\{\cdot \parallel \cdot\}$ is the Kullback-Leibler (KL) divergence between two probability distributions. The minimization of the KL-divergence is equivalent to the maximization of the free energy, since the first term of Eq. (8) does not depend on the trial posterior distribution Q. When the trial posterior distribution is equivalent to the true posterior distribution, the free energy is equivalent to the log marginal likelihood, $\ln P(Y)$.

In the VB method, we assume a factorized trial posterior distribution $Q(\Theta, X, Z) = Q(\Theta)Q(X, Z)$. Under this assumption, the maximization of the free energy is implemented as an efficient iterative algorithm similar to the EM algorithm. The details of the estimation algorithm for the CGM are provided in Appendix. After the algorithm converges, the free energy, which approximates the log marginal likelihood, can be used for determining the model structure, i.e., the cluster number M, because $\ln P(Y) = \ln P(Y|M)$ is the log likelihood of M.

3. Meta-Cluster Analysis

In many clustering methods, such as the k-means and model-based clustering methods, clustering results depend on the initial condition of the algorithm. Our clustering method based on the Bayes inference involves an integration over the parameters, and the dependence on the parameter initialization is removed, in principle. However, it still depends on the initial set-up for the parameters of the trial posterior distribution in the VB method. In order to cope with this problem, we propose a meta-cluster analysis procedure. First, we run a clustering algorithm many times with various initial conditions and take C good results from them. The goodness of each result is evaluated by an appropriate criterion, the log marginal likelihood $\ln P(Y)$ in our method. Then, we calculate the averaged similarity $h(i, j)$ between the i-th and j-th expression patterns using the C clustering results:

$$h(i, j) \equiv \frac{1}{C} \sum_{c=1}^{C} \sum_{m=1}^{M_c} P_c(z_m(i) = 1|\boldsymbol{y}(i)) P_c(z_m(j) = 1|\boldsymbol{y}(j)), \tag{9}$$

where c indexes a clustering result and M_c is the number of components in the c-th clustering result. $P_c(z_m = 1|\boldsymbol{y}(i))$ denotes the marginalized posterior distribution (Eq. (7)) of the c-th clustering result.

Our meta-cluster analysis is performed such that the following objective function is maximized:

$$E = \sum_{\tilde{m}=1}^{\tilde{M}} \sum_{i \in I_{\tilde{m}}} \sum_{j \in I_{\tilde{m}}} h(i, j) + \sum_{\tilde{m}_1=1}^{\tilde{M}} \sum_{\tilde{m}_2 \neq \tilde{m}_1}^{\tilde{M}} \sum_{i \in I_{\tilde{m}_1}} \sum_{j \in I_{\tilde{m}_2}} (1 - h(i, j)), \tag{10}$$

where $\tilde{M}$ is the number of meta clusters. I_m denotes the index set of genes in the m-th meta-cluster. Equation (10) is the sum of the internal similarity within each meta cluster (the first term) and the external dissimilarity between all meta cluster pairs (the second term). The validity of this objective function is discussed in the Results section. We apply the k-means clustering method many times for the set of N similarity vectors

$$\boldsymbol{h}_i = (h(i, 1), h(i, 2), \ldots, h(i, N)) \ \ (i = 1, \ldots, N), \tag{11}$$

and choose a meta-clustering result that maximizes the objective function (10).

Because we do not have *a priori* information about the characters of the similarity vectors (11), the non-parametric k-means clustering is sufficient in this meta-clustering analysis.

4. Result

4.1 Data Description

Our clustering method was applied to a gene expression time-series dataset of *Bacillus subtilis* during sporulation measured by cDNA microarray technology. The expression levels of 4,010 genes of Bacillus subtilis were measured every 30 minutes (9 hours, 19 time points) during sporulation. We removed the first and second time points from the original dataset, because we considered that sporulation had not started at that time. Consequently, each gene expression pattern is represented by a 17-dimensional vector. We used 'GeneSpring' (Silicon Genetics), a software for analyzing gene expression data, in order to eliminate two types of biases involved in the gene expression data; the intensity dependent bias was removed by a nonlinear transformation, and the array dependent bias was corrected by the median of the expression levels. After that, missing values in the corrected expression patterns were imputed by the k-nearest neighbor method [35]. Expression patterns with four or more missing values were not used, because they were harmful to the cluster analysis. Finally, we selected 617 expression patterns whose average expression level over the 17 time points was greater than zero, because these activated genes were expected to be related to sporulation.

The sporulation process can biologically be divided into five stages. At each stage, some sigma-factors activate the expression of a specific set of genes [36]. On the basis of biological knowledge with regard to transcription factors, we further divided the five stages into ten stages, and 89 genes (see **Table 1**) whose each function during the sporulation is known were separated into ten groups corresponding to the ten stages. Namely, each of the 89 genes out of 617 has one of 10 labels. Each clustering was conducted by using the gene expression dataset including all the 617 genes and the results were evaluated by comparison with such biological partition with respect to the labeled 89 genes. We used the adjusted Rand index (ARI; in Ref. 37)) to evaluate the similarity of two partitions, i.e., a clustering result obtained by our method and the biological partition. A high ARI value means that the corresponding clustering result agrees well with the biological partition [10].

4.2 Free Energy and ARI

We show the cluster analysis results based on the CGM estimated by the VB method (VB-CGM). First, we examined the model's dependence on the number of clusters. We prepared CGM models with cluster numbers from 1 to 20. For each cluster number, 50 kinds of initial set-ups were randomly prepared, and the VB method was applied with each initial set-up. The free energy and ARI were calculated for each of the results. **Figure 2** (a-b) shows schematic plots of the free energy and ARI, respectively, versus the number of clusters. The median of the free energy is maximal at 7 clusters. On the other hand, the ARI becomes almost flat, especially with more than 8 clusters. For comparison, instead of the VB method, we used an ML estimation for the same GCM models (ML-GCM). When evaluating models estimated by an ML estimation method, the Bayesian information criterion (BIC) [32] is often used [11]. The BIC is maximal at 15 clusters, while the ARI is maximal at 9 clusters (Fig. 2 (c-d)). The ARI decreases as the number of clusters increases. These results show that the free energy in the VB method exhibits good agreement with the ARI, while in our model's case the BIC in the ML estimation behaves differently from the ARI. **Figure 3** shows schematic plots of the number of effective clusters in CGM models, whose mixing rate g_m exceeds $0.5/M$, versus the number of clusters. When the VB method is used (Fig. 3 (a)), the number of effective clusters is consistently within the range of 5 to 7, implying that there are effective clusters with such a number in the dataset. When the ML estimation is used (Fig. 3 (b)), on the other hand, the number of effective clusters is proportional to the number of clusters M in the model. This implies that the ML estimation divides a proper cluster into small portions when M increases. Such division degrades the quality of the clustering results. The VB method does not exhibit such an improper cluster division. Accordingly, the VB-GCM is more robust than the ML-GCM method against the increase in the cluster number; moreover, the free energy can be a criterion to determine the appropriate number of effective clusters in the dataset.

Table 1 Partition of 89 known genes into 10 groups based on their known transcription factors. The numbers next to the gene names denote the meta-cluster indices obtained by our method.

Transcription factor	Gene				
sigH	spoVG(1)	spoVS(1)			
sigH & spo0A	sigF(2)	spo0A(1)	spo0F(1)	spoIIAA(2)	spoIIAB(2)
sigA & spo0A	sigE(2)	spoIIB(2)	spoIIE(2)	spoIIGA(2)	
sigE	cwlJ(5)	dacB(2)	mmgA(4)	mmgB(4)	mmgC(5)
	mmgD(4)	mmgE(4)	nucB(3)	spoIIIAA(4)	spoIIIAC(4)
	spoIIIAE(4)	spoIIIAG(3)	spoIIIAH(4)	spoIIIC(5)	spoIIP(4)
	spoIVA(4)	spoIVCB(4)	spoVID(4)	spoVR(4)	yisO(5)
	yrbA(5)	yrbB(5)	ysxE(4)		
sigE & spoIIID	cotJA(5)	cotJB(4)	cotJC(4)	spoIIID(5)	spoIVCA(4)
sigF	sigG(3)	spoIIQ(4)	sspF(6)		
sigF & sigG	dacF(5)	gpr(4)			
sigG	gerBA(5)	gerBC(4)	sleB(5)	splB(2)	spoVAB(5)
	spoVAC(5)	spoVAD(6)	spoVAE(6)	spoVAF(5)	sspB(5)
	ycxE(5)	ypeB(5)			
sigK	cotT(6)	spoIVFA(4)	spsA(7)	spsB(7)	spsC(7)
	spsD(7)	spsE(7)	spsF(7)	spsG(7)	spsI(7)
	spsJ(6)	spsK(6)	yisC(4)	yisD(4)	yisE(4)
	yisF(3)	yisG(5)	yjmC(4)	yjmD(4)	yjmF(4)
	yjmG(3)				
sigK & gerE	cgeB(7)	cgeD(2)	cgeE(2)	cotC(6)	cotD(7)
	cotG(7)	cotS(6)	cotV(7)	cotW(7)	cotX(7)
	cotY(6)	cotZ(6)			

4.3 Results of Meta-Cluster Analysis

Meta-cluster analysis tries to obtain a robust clustering result by integrating various clustering results after changing the initial set-up of the target clustering algorithm. In the meta-cluster analysis, we calculated $h(i, j)$ for 50 CGM clustering results, each of which has 7 clusters; this cluster number was determined by the free energy criterion (see Fig. 2 (a)). Consistent with this setting, we also set the number of meta clusters to 7. For comparison, we applied the same meta-clustering procedure to the results obtained by the k-means clustering with the Pearson's correlation coefficient. The meta-cluster analysis is formulated as optimization of the objective function (10).

The validity of this objective function (10) is checked here by comparing its values with the ARI values for 1,000 meta-clustering results starting from random initial conditions in the meta-clustering procedure. The objective function exhibits positive correlation with the ARI in both cases, by the VB-CGM and by the k-means clustering (**Fig. 4**). This implies that the objective function (10) can be used for choosing good results from various meta-clustering results. Figure 4

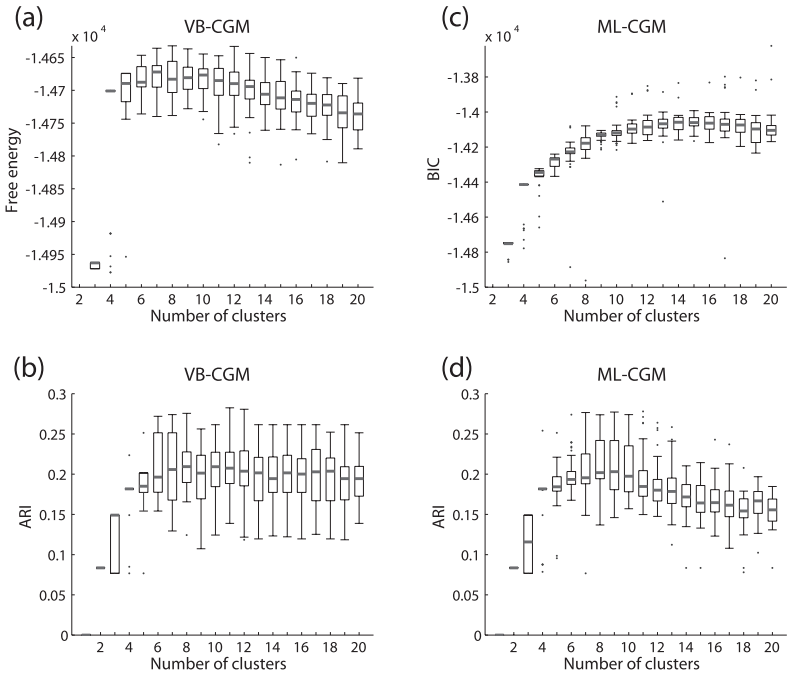


Fig. 2 Clustering results by the GCM model. (a) Schematic plot of the variational free energy when applying the VB method to the CGM models (VB-CGM) with various numbers of clusters. Horizontal and vertical axes denote the number of clusters and the free energy, respectively. Bold lines in the schematic plots denote the median. (b) The corresponding ARI. (c) Schematic plot of the BIC when applying an ML estimation to the CGM models (ML-CGM) with various numbers of clusters. (d) The corresponding ARI.

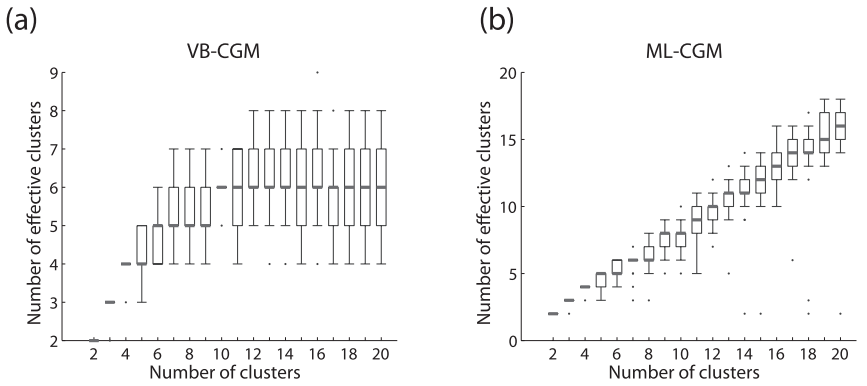


Fig. 3 The number of effective clusters in the CGM models. (a) Schematic plot of the number of effective clusters versus the number of all clusters by the VB-CGM. Note that the vertical axis represents discrete values. (b) Schematic plot of the number of effective clusters versus the number of all clusters by the ML-CGM.

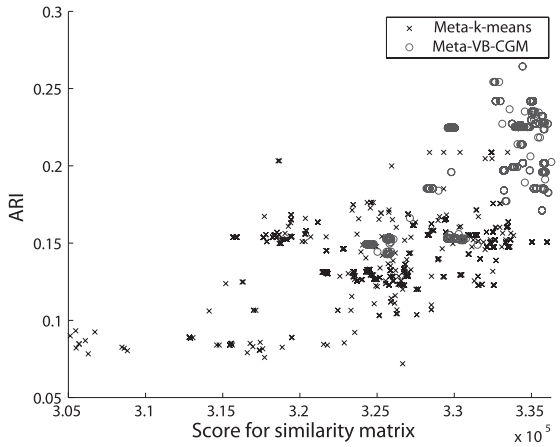


Fig. 4 Objective function of meta-cluster analysis and ARI. Scatter plot of the ARI versus the objective function (Eq. (10)), in the VB-CGM method (circles) and the k-means clustering (crosses).

also shows that the meta-clustering results based on the VB-CGM have larger ARI values than those based on the k-means clustering.

We further examined the meta-clustering result with the largest objective function value. In order to extract representative expression patterns from the meta-clustering result, we applied the probabilistic model (2) to the expression patterns in each of the meta clusters. The result is shown in **Fig. 5**. The representative expression patterns indicate that genes in the seven meta clusters were activated in turn during the sporulation process (right panels). It should be noted that the meta clusters, especially the second and fourth meta clusters, consist of expression patterns with various magnitudes (left panels).

Table 1 shows the partition of the 89 biologically known genes on the basis of

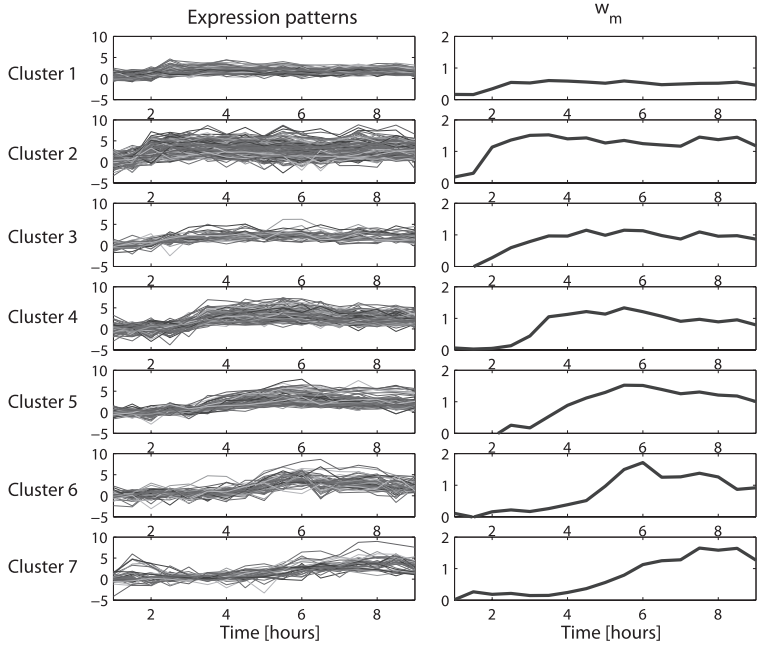


Fig. 5 Gene expression patterns partitioned by meta-cluster analysis. Each row corresponds to each of the seven meta clusters. The left and right panels represent expression patterns of the constituent genes and the representative expression pattern, $\boldsymbol{w}$ in the probabilistic model, respectively, in each of the seven meta clusters.

the meta-cluster analysis. The first and second meta clusters include genes that are known to be activated in the early stages of the sporulation process (sigH; sigH & spo0A; sigA & spo0A). In particular, the second meta cluster does not include genes activated in the middle or later stages, though the meta cluster consists of expression patterns of various magnitudes. The fourth and fifth meta clusters correspond to the middle stages (sigE; sigE & spoIIID; sigF; sigF& sigG). Genes in the sixth and seventh meta clusters were activated in the later stages (sigG; sigK; sigK & gerE).

4.4 Performance Comparison of Clustering Methods

For comparing the performances between the proposed method and the conventional clustering methods, we applied the ordinary Gaussian mixture (GM) model with the ML estimation (ML-GM), k-means clustering and hierarchical clustering to the same dataset. Since the setting of full covariance makes the ML estimation unstable, the diagonal covariance structure was employed in the GM model. We used the Pearson's correlation coefficient as the similarity measure in the k-means and the hierarchical clustering.

Table 2 Clustering method and the resulting ARI. Median of ARIs and the median absolute deviation for the optimal cluster number are shown for the upper four methods; maximal ARI is shown for three variations of hierarchical clustering.

Clustering method	ARI	Number of clusters	(Selection criterion)
VB-CGM	0.206 ± 0.040	7	(free energy)
ML-CGM	0.172 ± 0.014	15	(BIC)
Meta-VB-CGM	0.225 ± 0.017	7	(free energy)
ML-GM	0.150 ± 0.018	10	(BIC)
k-means	0.167 ± 0.022	10	(best choice)
Hierarchical (Average linkage)	0.177	142	(best choice)
Hierarchical (Complete linkage)	0.207	244	(best choice)
Hierarchical (Single linkage)	0.072	58	(best choice)

Table 2 summarizes the performance measure (ARI) of the eight clustering methods: VB-CGM, ML-CGM, meta-cluster analysis with VB-CGM (Meta-VB-CGM), ML-GM, k-means, and three variations of hierarchical clustering (average linkage, complete linkage, and single linkage). In VB-CGM, ML-CGM, Meta-VB-CGM, and ML-GM, we selected the cluster number based on their own criteria and calculated median and median absolute deviation of the ARIs. Meanwhile, in other methods, we heuristically selected the cluster number so that the median ARI (k-means) or ARI (hierarchical clusterings) show the best result. Note that selecting the best result is not realistic in a typical unsupervised situation. This result clearly shows that the VB-CGM enables us to automatically and objectively obtain clustering results that are consistent with biological knowledge. The result also indicates that the meta-cluster analysis (Meta-VB-CGM) can enhance not only the stability but also the clustering performance of the VB-CGM.

The performance of ML-GM was worse than those of ML-CGM and k-means. This result represents that Euclidean or Mahalanobis type distance between gene expression profiles is less accurate while correlation-based similarity measure can more appropriately capture co-expressed structures in the dataset.

Accordingly, our meta-cluster analysis incorporating the VB-CGM clustering is effective in extracting characteristic time-series while appropriately being insensitive to the magnitude, i.e., a robust correlation-based cluster analysis.

5. Discussion and Conclusions

Our newly-proposed CGM model assumes that genes in the same cluster are activated simultaneously and that their magnitudes obey a normal distribution. Observed expression patterns can be biased depending on the extent to which the assumption is incorrect. Our assumption can be verified by means of the distribution of the bias, which can be estimated using the probabilistic model of the observation noise process. Although the bias and the noise are not distinguished in our probabilistic model, discrimination can be achieved by assuming a more precise model.

The CGM model does not incorporate the time structure of the gene expression datasets explicitly, but the performance is comparable or better than that of recently proposed smooth spline clustering [16] which is a specialized method for time-course gene expression datasets (data not shown).

There have been studies made to improve the ability of Gaussian mixture models by introducing appropriate constraints (e.g., Ref. 13), 38)). The constraints in the CGM are stronger than those in the previous studies. The current results suggested that such constraints contributed to the stability of the clustering results. Moreover, the clustering results by the Bayesian inference (VB-CGM) were more stable than those by the ML estimation (ML-CGM).

Other approximation methods can be applied to the CGM model. The MCMC is the most accurate method, but is computationally expensive, and moreover it is difficult to check the convergence of the algorithm. On the other hand, the convergence of the VB algorithm can be easily checked by monitoring the free energy.

One of the simplest alternatives is the maximum *a posteriori* (MAP) estimation, in which the posterior of the model parameters is maximized. The MAP estimation works as a penalized maximum likelihood estimation [39] and is expected to avoid overfitting. Once the most probable parameters are obtained, the marginal likelihood is approximated based on the Taylor expansion of the posterior, i.e., Laplace approximation. The BIC is derived as the large sample limit for this Laplace approximation.

The VB method and the Laplace approximation were formerly compared on problems with low-dimensional parameter spaces [40]. That study concluded that the VB method is less accurate than the Laplace method. However, the VB method can be more stable than the Laplace method for high-dimensional parameter spaces; the latter requires computation of the determinant of the Hessian in the parameter space.

Another alternative is the evidence framework [30),41)], which is the same, in principle, as the model selection based on the marginal likelihood $\ln P(Y|M)$. The relationship between the evidence framework and the VB method was discussed in Ref. 41) for linear regression models. However, such a relationship has not been discussed for mixture models.

The meta-cluster analysis we proposed can be applied to an arbitrary clustering algorithm whose results are not unique due to the effects from initial set-up. Since poor clustering results degrade the meta-clustering result, however, we need a criterion to choose good clustering results that are appropriate the meta-cluster analysis. The free energy in the VB-CGM can be used for such a criterion.

In this study, we used the k-nearest neighbor method in order to estimate the missing values in the dataset. Probabilistic models, such as the CGM, can be extended to handle missing values [42]. Moreover, labeled data can be used to improve the quality of the clustering results [43], a technique that is termed semi-supervised. For our future work, we plan to introduce the above extensions and to assess the proposed method for various gene expression datasets.

References

1) Zhong, S. and Ghosh, J.: A Unified Framework for Model-based Clustering, *J. Mach. Learn. Res.*, Vol.4, pp.1001–1037 (2003).
2) Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D.: Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, Vol.95, No.25, pp.14863–14868 (1998).
3) Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M.: Systematic determination of genetic network architecture, *Nat. Genet.*, Vol.22, No.3, pp.281–285 (1999).

4) Kohonen, T.: *Self-Organizing Map*, 3rd edition, Springer-Verlag, Berlin (2001).
5) Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R.: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci. USA*, Vol.96, No.6, pp.2907–2912 (1999).
6) Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, Vol.286, No.5439, pp.531–537 (1999).
7) Somogyi, R.: Making sense of gene-expression data, *Pharmainformatics*, pp.17–24 (1999).
8) Dembélé, D. and Kastner, P.: Fuzzy C-means method for clustering microarray data, *Bioinformatics*, Vol.19, No.8, pp.9730–9780 (2003).
9) Tomida, S., Hanai, T., Honda, H. and Kobayashi, T.: Analysis of expression profile using fuzzy adaptive resonance theory, *Bioinformatics*, Vol.18, No.8, pp.1073–1083 (2002).
10) Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L.: Model-based clustering and data transformations for gene expression data, *Bioinformatics*, Vol.17, No.19, pp.977–987 (2001).
11) Fraley, C. and Raftery, A.E.: MCLUST: Software for model-based clustering, discriminant analysis and density stimation, Technical report, Technical Report No.415R, Department of Statistics, University of Washington (2002).
12) Ghosh, D. and Chinnaiyan, A.M.: Mixture modeling of gene expression data from microarray experiments, *Bioinformatics*, Vol.18, No.2, pp.275–286 (2002).
13) McLachlan, G.J., Bean, R.W. and Peel, D.: A mixture model-based approach to the clustering of microarray expression data, *Bioinformatics*, Vol.18, No.3, pp.413–422 (2002).
14) Banerjee, A., Dhillon, I.S., Ghosh, J. and Sra, S.: Clustering on the unit hypersphere using von Mises-Fisher distributions, *J. Mach. Learn. Res.*, Vol.6, pp.1345–1382 (2005).
15) Medvedovic, M. and Sivaganesan, S.: Bayesian infinite mixture model based clustering of gene expression profiles, *Bioinformatics*, Vol.18, No.9, pp.1194–1206 (2002).
16) Ma, P.C., Chan, K.C. and Chiu, D.K.: Clustering and re-clustering for pattern discovery in gene expression data, *J. Bioinform. Comput. Biol.*, Vol.3, No.2, pp.2810–301 (2005).
17) Tseng, G.C. and Wong, W.H.: Tight clustering: A resampling-based approach for identifying stable and tight patterns in data, *Biometrics*, Vol.61, No.1, pp.10–16 (2005).
18) Kaufman, L. and Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York (1990).
19) Thalamuthu, A., Mukhopadhyay, I., Zheng, X. and Tseng, G.C.: Evaluation and comparison of gene clustering methods in microarray analysis, *Bioinformatics* (2006).
20) Attias, H.: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Proc. 15th Conference on Uncertainty in Artificial Intelligence*, pp.21–30 (1999).
21) Waterhouse, S., MacKay, D.J.C. and Robinson, T.: Bayesian methods for mixtures of experts, *Advances in Neural Information Processing Systems*, Vol.8, pp.351–357 (1996).
22) Hyvaerinen, A., Hoyer, P. and Oja, E.: Image denoising by sparse code shrinkage, *Intelligent signal processing*, S, H. and B, K. (eds.), pp.554–568, IEEE Press (2001).
23) Olshausen, B.A., Sallee, P. and Lewicki, M.S.: Learning sparse image codes using a wavelet pyramid architecture, *Advances in Neural Information Processing Systems*, Vol.13, pp.887–893 (2000).
24) Zibulevsky, M. and Pearlmutter, B.A.: Blind source separation by sparse decomposition in a signal dictionary, *Neural Comput.*, Vol.13, No.4, pp.863–882 (2001).
25) Chen, Y., Dougherty, E.R. and Bittner, M.L.: Ratio-based decisions and the quantitative analysis of cDNA microarray images, *J. Biomedical Optics*, Vol.2, No.4, pp.364–374 (1997).
26) Held, G.A., Grinstein, G. and Tu, Y.: Modeling of DNA microarray data by using physical properties of hybridization, *Proc. Natl. Acad. Sci. USA*, Vol.100, No.13, pp.7575–7580 (2003).
27) Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W.: On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, *J. Comput. Biol.*, Vol.8, No.1, pp.37–52 (2001).
28) Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm, *J. Roy Statistical Society B*, Vol.39, No.1, pp.1–38 (1977).
29) Bishop, C.M.: *Neural networks for pattern recognition*, Oxford University Press, New York (1995).
30) MacKay, D.J.C.: A practical Bayesian framework for Backprop networks, *Neural Comput.*, Vol.4, pp.448–472 (1992).
31) Akaike, H.: A new look at the statistical model identification, *IEEE transactions on automatic control*, Vol.19, No.6, pp.716–723 (1974).
32) Schwarz, G.: Estimating the dimension of a model, *Annals of Statistics*, Vol.6, pp.461–464 (1978).
33) Liu, J.S.: *Monte carlo strategies in scientific computing*, Springer-Verlag, New York (2001).
34) Kass, R.E. and Raftery, A.E.: Bayes factors, *J. Am. Stat. Assoc.*, Vol.90, pp.773–795 (1995).

35) Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B.: Missing value estimation methods for DNA microarrays, *Bioinformatics*, Vol.17, No.6, pp.520–525 (2001).
36) Haldenwang, W.G.: The sigma factors of *Bacillus subtilis*, *Microbiol. Rev.*, Vol.59, pp.1–30 (1995).
37) Hubert, L.J. and Arabie, P.: Comparing partitions, *J. Classif.*, Vol.2, pp.193–218 (1985).
38) Tipping, M.E. and Bishop, C.M.: Mixtures of probabilistic principal component analyzers, *Neural Comput.*, Vol.11, No.2, pp.443–482 (1999).
39) Ormoneit, D. and Tresp, V.: Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates, *IEEE Trans. Neu. Net.*, Vol.9, No.4, pp.639–650 (1998).
40) Minka, T.P.: Using lower bound approximate integrals (2001).
41) MacKay, D.J.C.: Comparison of approximate methods for handling hyperparameters, *Neural Comput.*, Vol.11, pp.1035–1068 (1999).
42) Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S.: A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, Vol.19, No.16, pp.2088–2096 (2003).
43) Nigam, K., McCallum, A., Thrun, S. and Mitchell, T.: Text classification from labeled and unlabeled documents using EM, *Machine Learning*, Vol.39, No.2, pp.103–134 (2000).

Appendix

A.1 Component Model

Let $x(i)$ be a hidden random variable that obeys a normal distribution whose mean and variance are μ and 1, respectively. A D-dimensional observation vector $\boldsymbol{y}(i)$ for the i-th datum is given by a linear transformation of $x(i)$ with an additive noise:

$$\boldsymbol{y}(i) = \boldsymbol{w}x(i) + \boldsymbol{\epsilon}(i), \tag{12}$$

where $\boldsymbol{\epsilon}(i)$ is a D-dimensional noise vector that obeys $\mathcal{N}(\boldsymbol{\epsilon}(i)|0, \sigma \boldsymbol{I}_D)$. $\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \Sigma)$ denotes a multivariate normal distribution defined by

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \Sigma) \equiv \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})'\Sigma(\boldsymbol{x}-\boldsymbol{m}) + \frac{1}{2}\ln|\Sigma| - \frac{d}{2}\ln(2\pi)\right], \tag{13}$$

where $\boldsymbol{m}$ and Σ denote the mean and inverse covariance matrix, respectively. $\boldsymbol{I}_D$ denotes a D-by-D identity matrix. Let θ be the parameter set of the component model, i.e., $\theta \equiv \{\boldsymbol{w}, \mu, \sigma\}$. The complete data set consists of the observation data $Y \equiv \{\boldsymbol{y}(i)\}_{i=1}^N$ and hidden variables $X \equiv \{x(i)\}_{i=1}^N$. The likelihood of the complete data set is

$$P(Y, X|\theta) = \prod_{i=1}^{N} P(\boldsymbol{y}(i)|x(i), \boldsymbol{w}, \sigma)P(x(i)|\mu). \tag{14}$$

The likelihood of the observed data Y is given by integrating out the hidden variables X in Eq. (14): $P(Y|\theta) = \int dX P(Y, X|\theta)$.

A.2 Variational Bayes Method for the Component Model

A.2.1 General Theory

In the Bayesian framework, the posterior distribution of unknown variables, hidden variables and model parameters, is obtained according to the Bayes theorem:

$$P(X, \theta|Y) = \frac{P(Y, X|\theta)P_0(\theta)}{P(Y)}, \tag{15}$$

where $P_0(\theta)$ is the prior distribution of the model parameters and $P(Y)$ is the marginal likelihood. We prepare a trial posterior distribution $Q(X, \theta)$ that approximates the true posterior distribution $P(X, \theta|Y)$, and consider a (variational) free energy function defined by

$$\begin{aligned}\mathcal{F} &\equiv \int dX d\theta Q(X, \theta) \ln \frac{P(Y, X|\theta)P_0(\theta)}{Q(X, \theta)} \\ &= \ln P(Y) - KL\left[Q(X, \theta) \parallel P(X, \theta|Y)\right],\end{aligned} \tag{16}$$

where

$$\mathrm{KL}[Q(x) \parallel P(x)] = \int dx Q(x) \ln \frac{Q(x)}{P(x)} \tag{17}$$

is the Kullback-Leibler divergence. Equality (16) holds for an arbitrary trial posterior Q, showing that the true posterior is obtained by maximizing the free energy with respect to the trial posterior.

A variational Bayes method assumes a factorized trial posterior $Q(X, \theta) = Q(X)Q(\theta)$. Although this restriction introduces a bias in approximating the true posterior in general, the maximization of the free energy becomes tractable, as shown later. In order to maximize the free energy with respect to the trial posterior $Q(X)$, we consider a variational problem $\delta F/\delta Q(X) = 0$, which is solved as

$$\ln Q(X) = \langle \ln P(Y, X, \theta) \rangle_\theta + const., \tag{18}$$

where $\langle f(X,\theta)\rangle_\theta = \int d\theta Q(\theta) f(X,\theta)$. Similarly, $\delta F/\delta Q(\theta) = 0$ is solved as

$$\ln Q(\theta) = \langle \ln P(Y,X,\theta)\rangle_X + const.. \tag{19}$$

The free energy is maximized by alternately updating trial posteriors, $Q(X)$ and $Q(\theta)$, which is similar to the EM algorithm in a maximum likelihood estimation.

A.2.2 Parameter Estimation for the Component Model

When the conjugate prior is introduced into the component model, the probability of the complete data set $\{Y,X\}$ and the model parameters θ is given by

$$\begin{aligned} P(Y,X|\theta)P_0(\theta) &= P(Y,X,\boldsymbol{w},\sigma,\mu) \\ &= \prod_{i=1}^{N} \left[P(\boldsymbol{y}(i)|x(i),\boldsymbol{w},\sigma)P(x(i)|\mu)\right] P_0(\boldsymbol{w})P_0(\sigma)P_0(\mu) \end{aligned} \tag{20}$$

$$\begin{aligned} P_0(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|\bar{\boldsymbol{w}}_0, \alpha_0 \boldsymbol{I}_D) \\ P_0(\sigma) &= \mathcal{G}(\sigma|\gamma_{\sigma 0}, \bar{\sigma}_0) \\ P_0(\mu) &= \mathcal{N}(\mu|0, c_0). \end{aligned}$$

$\mathcal{G}(x|\gamma,m)$ denotes a gamma distribution

$$\mathcal{G}(x|\gamma,m) \equiv \exp\left[-\gamma m^{-1}x + (\gamma-1)\ln x + \gamma\ln(\gamma m^{-1}) - \ln\Gamma(\gamma)\right], \tag{21}$$

whose mean and variance are m and $\gamma^{-1}m^2$, respectively. According to Eq. (18) and (19), trial posterior distributions are obtained as follows:

$Q(X)$

$$\begin{aligned} \ln Q(X) &= \langle \ln P(Y,X,\boldsymbol{w},\sigma,\mu)\rangle_{\boldsymbol{w},\sigma,\mu} + const. \\ &= \left\langle \sum_{i=1}^{N}\left[-\frac{\sigma}{2}\parallel \boldsymbol{y}(i) - \boldsymbol{w}x(i)\parallel^2 - \frac{1}{2}(x(i)-\mu)^2\right]\right\rangle_{\boldsymbol{w},\sigma,\mu} + const. \\ &= \sum_{i=1}^{N}\left[-\frac{1}{2}\left\{(\bar{\sigma}\langle\parallel \boldsymbol{w}\parallel^2\rangle + 1)x^2(i) - 2(\boldsymbol{y}'(i)\bar{\sigma}\bar{\boldsymbol{w}} + \bar{\mu})x(i)\right\}\right] + const. \\ &= \sum_{i=1}^{N}\ln\mathcal{N}(x(i)|\bar{x}(i),\sigma_x) + const. \end{aligned} \tag{22}$$

$$Q(X) = \prod_{i=1}^{N}\mathcal{N}(x(i)|\bar{x}(i),\sigma_x), \tag{23}$$

where $\sigma_x = (\bar{\sigma}\langle\parallel \boldsymbol{w}\parallel^2\rangle + 1)$ and $\bar{x}(i) = \sigma_x^{-1}(\bar{\sigma}\bar{\boldsymbol{w}}'\boldsymbol{y}(i) + \bar{\mu})$. The average sufficient statistics are calculated as

$$E[x\boldsymbol{y}] = \frac{1}{N}\sum_{i=1}^{N}\bar{x}(i)\boldsymbol{y}(i) \tag{24}$$

$$E[\parallel \boldsymbol{y}\parallel^2] = \frac{1}{N}\sum_{i=1}^{N}\parallel \boldsymbol{y}(i)\parallel^2 \tag{25}$$

$$E[x] = \frac{1}{N}\sum_{i=1}^{N}\bar{x}(i) \tag{26}$$

$$E[x^2] = \sigma_x^{-1} + \frac{1}{N}\sum_{i=1}^{N}\bar{x}^2(i). \tag{27}$$

$Q(\boldsymbol{w})$

$$\begin{aligned} \ln Q(\boldsymbol{w}) &= \langle \ln P(Y,X,\boldsymbol{w},\sigma,\mu)\rangle_{X,\sigma,\mu} + const. \\ &= \left\langle -\frac{\sigma}{2}\sum_{i=1}^{N}\parallel \boldsymbol{y}(i) - \boldsymbol{w}x(i)\parallel^2 \right\rangle_{X,\sigma} - \frac{\alpha_0}{2}\parallel \boldsymbol{w} - \bar{\boldsymbol{w}}_0\parallel^2 + const. \\ &= -\frac{1}{2}\left[(\alpha_0 + \bar{\sigma}NE[x^2])\parallel \boldsymbol{w}\parallel^2 - 2\boldsymbol{w}'(\bar{\sigma}NE[x\boldsymbol{y}] + \alpha_0\bar{\boldsymbol{w}}_0)\right] + const. \end{aligned}$$

$$Q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\bar{\boldsymbol{w}}, \alpha\boldsymbol{I}_D), \tag{28}$$

where $\alpha = \alpha_0 + \bar{\sigma}NE[x^2]$ and $\bar{\boldsymbol{w}} = \alpha^{-1}(\bar{\sigma}NE[x\boldsymbol{y}] + \alpha_0\bar{\boldsymbol{w}}_0)$.

$Q(\sigma)$

$$\begin{aligned} \ln Q(\sigma) &= \langle P(Y,X,\boldsymbol{w},\sigma,\mu)\rangle_{X,\boldsymbol{w},\mu} + const. \\ &= \left\langle -\frac{\sigma}{2}\sum_{i=1}^{N}\parallel \boldsymbol{y}(i) - \boldsymbol{w}x(i)\parallel^2 + \frac{D}{2}\ln\sigma \right\rangle_{X,\boldsymbol{w}} \\ &\quad - \gamma_{\sigma 0}\bar{\sigma}_0^{-1}\sigma + (\gamma_{\sigma 0}-1)\ln\sigma + const. \\ &= -\left[\frac{N}{2}\left(E[\parallel \boldsymbol{y}\parallel^2] - 2\bar{\boldsymbol{w}}'E[x\boldsymbol{y}] + E[x^2]\langle\parallel \boldsymbol{w}\parallel^2\rangle\right) + \gamma_{\sigma 0}\bar{\sigma}_0^{-1}\right]\sigma \\ &\quad + \left(\frac{DN}{2} + \gamma_{\sigma 0} - 1\right)\ln\sigma + const. \end{aligned}$$

$$Q(\sigma) = \mathcal{G}(\sigma|\gamma_\sigma, \bar{\sigma}), \tag{29}$$

where $\gamma_\sigma = DN/2 + \gamma_{\sigma 0}$ and

$$\bar{\sigma} = \gamma_\sigma \left[\frac{N}{2} \left(E[\| \boldsymbol{y} \|^2] - 2\bar{\boldsymbol{w}}' E[x\boldsymbol{y}] + E[x^2]\langle \| \boldsymbol{w} \|^2 \rangle \right) + \gamma_{\sigma 0} \bar{\sigma}_0^{-1} \right]^{-1}.$$

$Q(\mu)$

$$\begin{aligned}
\ln Q(\mu) &= \langle \ln P(Y, X, \boldsymbol{w}, \sigma, \mu) \rangle_{X, \boldsymbol{w}, \sigma} + const. \\
&= -\frac{N}{2}(E[x^2] - 2E[x]\mu + \mu^2) - \frac{c_0}{2}\mu^2 + const. \\
&= -\frac{c}{2}(\mu - \bar{\mu})^2 + const. \\
Q(\mu) &= \mathcal{N}(\mu | \bar{\mu}, c),
\end{aligned} \tag{30}$$

where $c = N + c_0$ and $\bar{\mu} = c^{-1} N E[x]$.

A.2.3 Calculation of the Free Energy

The free energy function is decomposed as follows:

$$\mathcal{F} = L + H_X + H_{\boldsymbol{w}} + H_\sigma + H_\mu. \tag{31}$$

L

$$\begin{aligned}
L &= \langle \ln P(Y | X, \boldsymbol{w}, \sigma) \rangle_{X, \boldsymbol{w}, \sigma} \\
&= \left\langle -\frac{N\sigma}{2} \left(E[\| \boldsymbol{y} \|^2] - 2E[x\boldsymbol{y}]'\boldsymbol{w} + E[x^2] \| \boldsymbol{w} \|^2 \right) + \frac{ND}{2} \ln(\sigma / 2\pi) \right\rangle_{\boldsymbol{w}, \sigma} \\
&= -\frac{N\bar{\sigma}}{2} \left(E[\| \boldsymbol{y} \|^2] - 2E[x\boldsymbol{y}]'\bar{\boldsymbol{w}} + E[x^2]\langle \| \bar{\boldsymbol{w}} \|^2 \rangle_{\boldsymbol{w}} \right) \\
&\quad + \frac{ND}{2} \left(\langle \ln \sigma \rangle_\sigma - \ln 2\pi \right),
\end{aligned} \tag{32}$$

where $\langle \ln \sigma \rangle_\sigma = \ln \bar{\sigma} + \psi(\gamma_\sigma) - \ln \gamma_\sigma$. $\psi(x)$ is a digamma function.

H_X

$$\begin{aligned}
H_X &= \langle \ln[P(X|\mu)/Q(X)] \rangle_{X, \mu} \\
&= -\frac{N}{2}(E[x^2] - 2E[x]\bar{\mu} + \langle \mu^2 \rangle) + \frac{N}{2} \ln(1/2\pi) \\
&\quad + \frac{\sigma_x}{2} \sum_{i=1}^{N} (\langle x^2(i) \rangle_{x(i)} - 2\bar{x}^2(i) + \bar{x}^2(i)) - \frac{N}{2} \ln(\sigma_x / 2\pi) \\
&= -\frac{N}{2}(E[x^2] - 2E[x]\bar{\mu} + \langle \mu^2 \rangle) - \frac{N}{2} \ln \sigma_x + \frac{N}{2}.
\end{aligned} \tag{33}$$

$H_{\boldsymbol{w}}$

$$\begin{aligned}
H_{\boldsymbol{w}} &= \langle \ln[P_0(\boldsymbol{w})/Q(\boldsymbol{w})] \rangle_{\boldsymbol{w}} \\
&= -\frac{1}{2} \left[D \ln \alpha \alpha_0^{-1} + D\alpha^{-1}\alpha_0 - D + \alpha_0 \| \bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}_0 \|^2 \right].
\end{aligned} \tag{34}$$

H_σ

$$\begin{aligned}
H_\sigma &= \langle \ln[P_0(\sigma)/Q(\sigma)] \rangle_\sigma \\
&= \gamma_{\sigma 0} \left[\langle \ln \sigma \rangle_\sigma - \ln \bar{\sigma}_0 - \bar{\sigma}\bar{\sigma}_0^{-1} + 1 \right] + \Phi(\gamma_\sigma, \gamma_{\sigma 0}),
\end{aligned} \tag{35}$$

where $\Phi(\gamma, \gamma_0)$ is defined by

$$\Phi(\gamma, \gamma_0) = [\ln \Gamma(\gamma) - \gamma\psi(\gamma) + \gamma] - [\ln \gamma(\gamma_0) - \gamma_0 \ln \gamma_0 + \gamma_0]. \tag{36}$$

H_μ

$$\begin{aligned}
H_\mu &= \langle \ln[P_0(\mu)/Q(\mu)] \rangle_\mu \\
&= -\frac{1}{2} \left[\ln c c_0^{-1} + c^{-1}c_0 - 1 + c_0 \bar{\mu}^2 \right].
\end{aligned} \tag{37}$$

A.3 Constrained Gaussian Mixture Model

We consider a mixture of the component models:

$$P(\boldsymbol{y}(i), \boldsymbol{x}(i), \boldsymbol{z}(i) | \Theta) = \prod_{m=1}^{M} [P(\boldsymbol{y}(i), x_m(i) | \theta_m) g_m]^{z_m(i)}, \tag{38}$$

where M is the number of components, $\boldsymbol{x}(i) \equiv (x_1(i), \ldots, x_M(i))$ is the set of hidden variables, and $\boldsymbol{z}(i) \equiv (z_1(i), \ldots, z_M(i))$ is the set of indicator variables that satisfy $z_m(i) \in \{0, 1\}$ and $\sum_{m=1}^{M} z_m(i) = 1$. $\Theta \equiv \{\{\theta_m\}_{m=1}^{M}, \boldsymbol{g}\}$ is the set of model parameters. θ_m is the parameter set of the m-th component model and $\boldsymbol{g} = (g_1, \ldots, g_M)$ is the set of the mixing rate parameters that satisfies $g_m \geq 0$ and $\sum_{m=1}^{M} g_m = 1$. The mixture model (38) is called a constrained Gaussian mixture (CGM) model.

For the CGM model, the complete data set is given by $\{Y, X, Z\}$, where $Y \equiv \{\boldsymbol{y}(i)\}_{i=1}^{N}$, $X \equiv \{\boldsymbol{x}(i)\}_{i=1}^{N}$ and $Z \equiv \{\boldsymbol{z}(i)\}_{i=1}^{N}$. The joint probability of all variables is given by

$$P(Y, X, Z, \Theta) = \prod_{m=1}^{M} \left[\prod_{i=1}^{N} \{P(\boldsymbol{y}(i) | x_m(i), \boldsymbol{w}_m, \sigma_m) P(x_m(i) | \mu_m) g_m\}^{z_m(i)} P_0(\theta_m) \right] P_0(\boldsymbol{g}). \tag{39}$$

As in the general theory above, we assume a factorized trial posterior,

$Q(X,Z)Q(\boldsymbol{g})\prod_{m=1}^{M}Q(\theta_m)$, and a conjugate prior distribution

$$P_0(\Theta)=P_0(\boldsymbol{g})\prod_{m=1}^{M}P_0(\theta_m)$$
$$P_0(\boldsymbol{g})=\mathcal{D}(\boldsymbol{g}|\boldsymbol{\gamma}_0)$$
$$P_0(\theta_m)=P_0(\boldsymbol{w}_m)P_0(\sigma_m)P_0(\mu_m),$$

where $\boldsymbol{\gamma}_0=(\gamma_0^1,\ldots,\gamma_0^M)$ and $\mathcal{D}(\boldsymbol{g}|\boldsymbol{\gamma})$ denotes a Dirichlet distribution

$$\mathcal{D}(\boldsymbol{g}|\boldsymbol{\gamma})\equiv\exp\left[\sum_{m=1}^{M}\{\gamma^m\ln g_m-\ln\Gamma(\gamma^m+1)\}+\ln\Gamma\left(\sum_{n=1}^{M}\gamma^n+M\right)\right]. \tag{40}$$

The free energy function for the CGM model is then given by

$$\begin{aligned}\mathcal{F}=&\sum_{m=1}^{M}\left[\sum_{i=1}^{N}z_m(i)\left\{\langle\ln P(\boldsymbol{y}(i),x_m(i)|\theta_m)+\ln g_m-\ln Q(x_m(i),z_m(i)=1)\rangle\right\}\right.\\&+\langle[\ln P_0(\boldsymbol{w}_m)/Q(\boldsymbol{w}_m)]\rangle+\langle[\ln P_0(\sigma_m)/Q(\sigma_m)]\rangle+\left.\langle[\ln P_0(\mu_m)/Q(\mu_m)]\rangle\right]\\&+\langle\ln[P_0(\boldsymbol{g})/Q(\boldsymbol{g})]\rangle.\end{aligned} \tag{41}$$

A.3.1 Trial Distribution

The trial distribution that maximizes the free energy is obtained as follows.

$Q(X,Z)$

$$\begin{aligned}\ln Q(X,Z)&=\langle\ln P(Y,X,Z,\Theta)\rangle_\Theta+const.\\&=\sum_{i=1}^{N}\sum_{m=1}^{M}z_m(i)\left\{\langle\ln P(\boldsymbol{y}(i),x_m(i)|\boldsymbol{w}_m,\sigma_m,\mu_m)\rangle_{\boldsymbol{w}_m,\sigma_m,\mu_m}+\langle\ln g_m\rangle\right\}+const.\\&=\sum_{i=1}^{N}\ln Q(\boldsymbol{x}(i),\boldsymbol{z}(i)),\end{aligned} \tag{42}$$

where $\langle\ln g_m\rangle=(\gamma^m+1)/(\sum_{n=1}^{M}\gamma^n+M)$. $Q(\boldsymbol{x},\boldsymbol{z})$ can be decomposed as $Q(\boldsymbol{x},\boldsymbol{z})=Q(\boldsymbol{x}|\boldsymbol{z})Q(\boldsymbol{z})$. If $z_m=1$ and $z_n=0$ for $n\neq m$, $Q(x_m|\boldsymbol{z})$ is calculated by Eq. (23). $Q(z_m=1)$ is obtained by integrating out x_m:

$$\begin{aligned}Q(z_m=1)&\propto\int dx_m\exp\left[\langle\ln P(\boldsymbol{y},x_m|\boldsymbol{w}_m,\sigma_m,\mu_m)\rangle_{\boldsymbol{w}_m,\sigma_m,\mu_m}+\langle\ln g_m\rangle\right]\\&\equiv U_m\end{aligned} \tag{43}$$

$$\begin{aligned}U_m=&\int dx_m\exp\left[\left\langle-\frac{\sigma_m}{2}\parallel\boldsymbol{y}-\boldsymbol{w}_mx_m\parallel^2\right\rangle_{\boldsymbol{w}_m,\sigma_m}\right.\\&\left.+\frac{D}{2}\langle\ln(\sigma_m/2\pi)\rangle-\frac{1}{2}\langle(x_m-\mu_m)^2\rangle_{\mu_m}+\frac{1}{2}\ln(1/2\pi)+\langle\ln g_m\rangle\right]\\=&\int dx_m\exp\left[-\frac{1}{2}\left\{(\bar\sigma_m\langle\parallel\boldsymbol{w}_m\parallel^2\rangle+1)x_m^2-2(\bar\sigma_m\boldsymbol{y}'\bar{\boldsymbol{w}}_m+\bar\mu_m)x_m\right\}\right.\\&\left.-\frac{1}{2}\bar\sigma_m\parallel\boldsymbol{y}\parallel^2-\frac{1}{2}\langle\mu_m^2\rangle+\frac{D}{2}\langle\ln\sigma_m\rangle-\frac{D+1}{2}\ln(2\pi)+\langle\ln g_m\rangle\right]\\=&\int dx_m\exp\left[-\frac{\sigma_{xm}}{2}(x_m-\bar x_m)^2+\frac{\sigma_{xm}}{2}\bar x_m^2-\frac{1}{2}\bar\sigma_m\parallel\boldsymbol{y}\parallel^2-\frac{1}{2}\langle\mu_m^2\rangle\right.\\&\left.+\frac{D}{2}\langle\ln\sigma_m\rangle-\frac{D+1}{2}\ln(2\pi)+\langle\ln g_m\rangle\right]\\=&\exp\left[\frac{1}{2}\left\{\ln\sigma_{xm}^{-1}+\sigma_{xm}\bar x_m^2-\bar\sigma_m\parallel\boldsymbol{y}\parallel^2\right.\right.\\&\left.\left.-\langle\mu_m^2\rangle+D\langle\ln\sigma_m\rangle\ -D\ln(2\pi)\right\}+\langle\ln g_m\rangle\right],\end{aligned}$$

where $\sigma_{xm}=(\bar\sigma_m\langle\parallel\boldsymbol{w}\parallel^2\rangle+1)$ and $\bar x_m=\sigma_{xm}^{-1}(\bar\sigma_m\bar{\boldsymbol{w}}_m'\boldsymbol{y}+\bar\mu)$. $Q(z_m=1)$ is obtained by $Q(z_m=1)=U_m/\sum_{n=1}^{M}U_n$. The average sufficient statistics are calculated by

$$E_m[zx\boldsymbol{y}]=\frac{1}{N}\sum_{i=1}^{N}\bar z_m(i)\bar x_m(i)\boldsymbol{y}(i)$$

$$E_m[z\parallel\boldsymbol{y}\parallel^2]=\frac{1}{N}\sum_{i=1}^{N}\bar z_m(i)\parallel\boldsymbol{y}(i)\parallel^2$$

$$E_m[zx]=\frac{1}{N}\sum_{i=1}^{N}\bar z_m(i)\bar x_m(i)$$

$$E_m[zx^2] = \frac{1}{N}\sum_{i=1}^{N} \bar{z}_m(i)\{\sigma_{xm}^{-1} + \bar{x}_m^2(i)\}$$

$$E_m[z] = \frac{1}{N}\sum_{i=1}^{N} z_m(i).$$

$Q(\boldsymbol{w}_m)$

$$\ln Q(\boldsymbol{w}_m) = -\frac{1}{2}\left[(\alpha_{m0} + \bar{\sigma}_m N E_m[zx^2]) \parallel \boldsymbol{w}_m \parallel^2 \right.$$
$$\left. -2\boldsymbol{w}'(\bar{\sigma}_m N E_m[zx\boldsymbol{y}] + \alpha_{m0}\bar{\boldsymbol{w}}_{m0})\right] + const.$$
$$Q(\boldsymbol{w}_m) = \mathcal{N}(\boldsymbol{w}_m|\bar{\boldsymbol{w}}_m, \alpha_m \boldsymbol{I}_D), \tag{44}$$

where $\alpha_m = \alpha_{m0} + \bar{\sigma}_m N E_m[zx^2]$ and $\bar{\boldsymbol{w}}_m = \alpha_m^{-1}(\bar{\sigma}_m N E_m[zx\boldsymbol{y}] + \bar{\alpha}_{m0}\bar{\boldsymbol{w}}_{m0})$.

$Q(\sigma_m)$

$$\ln Q(\sigma_m) = -\left[\frac{N}{2}\left(E_m[z \parallel \boldsymbol{y} \parallel^2] - 2\bar{\boldsymbol{w}}' E_m[zx\boldsymbol{y}]\right.\right.$$
$$\left.\left. + E_m[zx^2]\langle \parallel \boldsymbol{w}_m \parallel^2\rangle\right) + \gamma_{\sigma m0}\bar{\sigma}_{m0}^{-1}\right]\sigma$$
$$+\left(\frac{DNE_m[z]}{2} + \gamma_{\sigma m0} - 1\right)\ln\sigma_m + const.$$
$$Q(\sigma_m) = \mathcal{G}(\sigma|\gamma_{\sigma m}, \bar{\sigma}_m), \tag{45}$$

where $\gamma_\sigma = DNE_m[z]/2 + \gamma_{\sigma m0}$ and

$$\bar{\sigma}_m = \gamma_{\sigma m}\left[\frac{N}{2}\left(E_m[z \parallel \boldsymbol{y} \parallel^2] - 2\bar{\boldsymbol{w}}' E_m[zx\boldsymbol{y}] + E_m[zx^2]\langle \parallel \boldsymbol{w} \parallel^2\rangle\right) + \gamma_{\sigma m0}\bar{\sigma}_{m0}^{-1}\right]^{-1}.$$

$Q(\mu_m)$

$$\ln Q(\mu_m) = -\frac{N}{2}(E_m[zx^2] - 2E_m[zx]\mu_m + E_m[z]\mu_m^2) - \frac{c_{m0}}{2}\mu_m^2 + const.$$
$$= -\frac{c_m}{2}(\mu_m - \bar{\mu}_m)^2 + const.$$
$$Q(\mu) = \mathcal{N}(\mu|\bar{\mu}_m, c_m), \tag{46}$$

where $c_m = NE_m[z] + c_0$ and $\bar{\mu} = c_m^{-1} N E_m[zx]$.

$Q(\boldsymbol{g})$

$$\ln Q(\boldsymbol{g}) = \langle \ln P(Y, X, Z, \Theta)\rangle_{X, Z, \{\theta_m\}_{m=1}^M} + const.$$
$$= \sum_{m=1}^{M} N(E_m[z] + \gamma_0^m)\ln g_m + const.$$
$$Q(\boldsymbol{g}) = \mathcal{D}(\boldsymbol{g}|\boldsymbol{\gamma}), \tag{47}$$

where $\boldsymbol{\gamma} = (\gamma^1, \ldots, \gamma^M)$ and $\gamma^m = \gamma_0^m + NE_m[z]$.

A.3.2 Calculation of the Free Energy

The free energy Eq. (41) is decomposed as

$$\mathcal{F} = \tilde{L} + H_g + \sum_{m=1}^{M}\{H_{\boldsymbol{w}m} + H_{\sigma m} + H_{\mu m}\}, \tag{48}$$

where

$$\tilde{L} = \sum_{i=1}^{N}\sum_{m=1}^{M}\frac{U_m(i)}{\sum_{n=1}^{M} U_n(i)}\langle[\ln P(\boldsymbol{y}(i), \boldsymbol{x}_m(i)|\boldsymbol{w}_m, \sigma_m, \mu_m)$$
$$+\ln g_m - \ln Q(x_m(i), z_m(i) = 1)]\rangle$$
$$= \sum_{i=1}^{N}\sum_{m=1}^{M}\frac{U_m(i)}{\sum_{n=1}^{M} U_n(i)}\ln\sum_{n=1}^{M} U_n(i) = \sum_{i=1}^{N}\ln\sum_{n=1}^{M} U_n(i). \tag{49}$$

Note that

$$Q(x_m(i), z_m(i) = 1) = \frac{\exp[\langle \ln P(\boldsymbol{y}(i), x_m(i)|\theta_m)\rangle_{\theta_m} + \langle \ln g_m\rangle]}{\sum_{n=1}^{M} U_n(i)},$$

then

$$\ln Q(x_m(i), z_m(i) = 1) = \langle \ln P(\boldsymbol{y}(i), x_m(i)|\theta_m)\rangle_{\theta_m} + \langle \ln g_m\rangle - \ln\sum_{n=1}^{M} U_n(i)$$

holds. $H_{\boldsymbol{w}m}$, $H_{\sigma m}$ and $H_{\mu m}$ have the same forms as Eq. (34), (35) and (37), respectively.

H_g

$$H_g = \langle \ln[P_0(\boldsymbol{g})/Q(\boldsymbol{g})]\rangle$$
$$= \sum_{m=1}^{M}\{(\gamma_0^m - \gamma^m)\langle \ln g_m\rangle - \ln\Gamma(\gamma_0^m + 1) + \ln\Gamma(\gamma^m + 1)\}$$

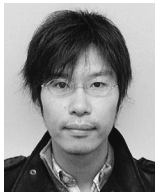
$$+\ln\Gamma\left[\sum_{m=1}^{M}\gamma_0^m+M\right]-\ln\Gamma\left[\sum_{m=1}^{M}\gamma^m+1\right], \tag{50}$$

where $\langle \ln g_m \rangle = \psi(\gamma^m + 1) - \psi\left(\sum_{n=1}^{M}\gamma^n + M\right)$.

(Received January 7, 2009)
(Accepted February 5, 2009)
(Released May 25, 2009)

(Communicated by *Yoichi Takenaka*)

Naoto Yukinawa was born in 1977. He received the B.S. degree in bioscience from Tokyo Institute of Technology in 2001. He received the Ph.D. degree in 2006 from Nara Institute of Science and Technology for studies on the statistical machine learning approaches for system identification and classification of gene expression profiles. Since 2006, he has worked as a researcher in the group of Dr. Shin Ishii at Kyoto University. His research interests include machine learning and their applications to transcriptomic and proteomic analyses.



Taku Yoshioka was born in 1975. He received the B.E. degree in 1997 from Osaka Electro Communication University, and M.E. and Ph.D. degrees in 1999 and 2001, respectively, from Nara Institute of Science and Technology. He is currently a researcher at Computational Neuroscience Laboratories, Advanced Telecommunication Research Institute. He has been interested in application of statistical learning method, and now he is working on solving the inverse problem for magnetoenchephalography.

Kazuo Kobayashi was born in 1968. He received the Ph.D. in 1997 from Tokyo University of Agriculture and Technology. Since 2000, he has worked as an assistance professor at Nara Institute of Science and Technology.

Naotake Ogasawara received Ph.D. from Nagoya University. In 1975, Research Associate in Cancer Research Institute, Kanazawa University. In 1986, Lecturer in Osaka University Medical School. In 1993, Professor in Graduate School of Biological Sciences, Nara Institute of Science and Technology. In 2002, Professor in Graduate School of Information Science, Nara Institute of Science and Technology. In 2007, Vice President of Nara Institute of Science and Technology. Since 2009, he has worked as a professor in Graduate School of Information Science, Nara Institute of Science and Technology.

Shin Ishii was born in 1962. He received the B.E., M.E. and Ph.D. degrees in 1986, 1988 and 1997, respectively, from the University of Tokyo. He is currently a professor at Graduate School of Informatics, Kyoto University, after a 10-year career at Graduate School of Information Science, Nara Institute of Science and Technology. He has been interested in statistical bioinformatics and systems neurobiology, and has approached these areas from both basic and practical viewpoints.