Modeling Cellular Processes with Variational Bayesian Cooperative Vector Quantizer

X. Lu, M. Hauskrecht, and R.S. Day

Pacific Symposium on Biocomputing 9:533-544(2004)

MODELING CELLULAR PROCESSES WITH VARIATIONAL BAYESIAN COOPERATIVE VECTOR QUANTIZER

X. LU[1,2,4], M. HAUSKRECHT[2] and R.S. DAY[3]
[1]Center for Biomedical Informatics, [2]Dept of Computer Science,
[3]Dept of Biostatistics. University of Pittsburgh
[4]Dept of Biometry and Epidemiology, Medical University of South Carolina
email: lux@musc.edu[a], milos@cs.pitt.edu, day@upci.pitt.edu

Abstract

Gene expression of a cell is controlled by sophisticated cellular processes. The capability of inferring the states of these cellular processes would provide insight into the mechanism of gene expression control system. In this paper, we propose and investigate the cooperative vector quantizer (CVQ) model for analysis of microarray data. The CVQ model could be capable of decomposing observed microarray data into many different regulatory subprocesses. To make the CVQ analysis tractable we develop and apply variational approximations. Bayesian model selection is employed in the model, so that the optimal number processes is determined purely from observed micro-array data. We test the model and algorithms on two datasets: (1) simulated gene-expression data and (2) real-world yeast cell-cycle microarray data. The results illustrate the ability of the CVQ approach to recover and characterize regulatory gene expression subprocesses, indicating a potential for advanced gene expression data analysis.

1 Introduction

Current DNA microarray technology allows scientists to monitor gene expression at genome level. Although microarray data are not direct measurements of activity of cellular processes (or signal transduction pathways), they provide opportunities to infer the states of the cellular processes and study the mechanism of gene expression control at the system level. When a cell is subjected to different conditions, the states of the processes controlling gene expression change accordingly and result in different gene expression patterns. One important task for system biologists is to identify the cellular processes controlling gene expression and infer their states under a specific condition based on observed expression patterns. Different approaches have been applied in order to identify the cellular processes by decomposing (deconvoluting) the observed microarray data into different components. For example, singular value decomposition (SVD)[1], principal component analysis (PCA)[2], independent component analysis (ICA)[3,4], Bayesian decomposition[5] and probabilistic

[a] To whom correspondence should be addressed.

relation modeling (PRM)[6] have been used to decompose observed microarray data into different processes.

The problem of identifying hidden regulatory processes in a cell can be formulated as a *blind source separation* problem, where distinct regulatory processes, which we would like to identify and characterize, are modeled as hidden sources[b]. The task is to identify the source signals purely based on observed data. An additional challenge is that the separation process must be performed fully unsupervised - the number of sources is not known in advance.

To facilitate biological interpretation, the originating signals of the processes in a system should be identified uniquely. Some of the aforementioned models, such as SVD and PCA, restrict the components to be orthonormal, thus they are not suitable for blind source separation. Independent component analysis (ICA), independent factor analysis (IFA) and various vector quantization models[7,8,9,10] are among the models used for blind source separation. In this work we develop an inference algorithm for one such model – the *cooperative vector quantizer* (CVQ) model. The main advantage of the CVQ model over other blind source separation models is that it mimics the switching-state nature of the regulatory processes; consequently, the results of the analysis can be easily interpreted by biologists.

Fully unsupervised blind source separation requires learning the model structure. In microarray data analysis, one needs to infer the optimal number of latent regulatory processes in the system. The parameters of a latent variable model with a fixed structure (known number of processes) can be learned using maximum likelihood estimation (MLE) techniques, e.g. the expectation maximization (EM)[11] algorithm, as in Segal et al[6]. Unfortunately, the value of likelihood by itself is not suitable for model selection. The main reason is that MLE prefers more complex models and tends to over-fit the training data. That is, more complex models return higher likelihood scores for the training data, but they do not generalize well to future, yet to be seen, data. On the other hand, the methods used in the studies by Alter et al[1] and Liebermeister[3] simply dictate the number of processes of the model and do not have the flexibility of model selection. Model selection can be addressed effectively within the Bayesian framework[12,13,14]. Bayesian selection penalizes models for complexity as well as for poor fit, therefore it implements Occam's Razor. In this work, we investigate the Bayesian model selection framework in the context of the CVQ model. More specifically, we derive and implement a variational Bayesian approach which can automatically learn both the structure and parameters of the CVQ model, and thus perform full-scale blind source separation.

In the following sections, we first present the CVQ model. After that, we discuss the theory of the Bayesian model selection and its approximations. We derive and present a variational Bayesian approximation for learning the CVQ model from data.

[b] We use "sources" and "processes" interchangeably throughout the rest of paper.

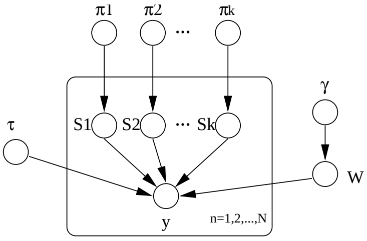

Figure 1: A directed acyclic graph (DAG) representation of the cooperative vector quantizer (CVQ) model. The square corresponds to an individual data point which consists of observed variables $\mathbf{y}$ and latent variables $\mathbf{s}$. $\mathbf{W}$, $\boldsymbol{\gamma}$, τ and $\boldsymbol{\pi}$ are model parameters.

Finally, we test the model and algorithms on (1) a simulated gene expression data (2) yeast cell-cycle microarray data[20] and discuss the results.

2 The CVQ Model

In the CVQ model, the states of the cellular processes are represented as a set of binary variables $\mathbf{s} = \{s_k\}_{k=1}^{K}$ referred to as sources, where K is the number of processes in a given model. Each source assumes a value of 0/1, which simulate the "off/on" state of cellular processes. Each microarray experiment is represented as a D-dimensional vector $\mathbf{y}$, where D is the number of genes on a microarray. An observed data point $\mathbf{y}^{(n)}$ is produced cooperatively by the sources depending on their states. When a source s_k equals 1, it will output a D-dimensional weight $\mathbf{w}_k$ to $\mathbf{y}$. We can think of the source variable s_k as a switch which, when turned on, allows the outflow of weights $\mathbf{w}_k$ to $\mathbf{y}$. More formally

$$\mathbf{y} = \sum_{k=1}^{K} s_k \mathbf{w}_k + \epsilon \qquad\qquad P(\mathbf{y}|\mathbf{s}) \sim \mathcal{N}\left(\mathbf{y}|\sum_{k=1}^{K} s_k \mathbf{w}_k, \boldsymbol{\Lambda}\right)$$

where $\mathcal{N}(.|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate Gaussian distribution; s_k is an index function; $\mathbf{w}_k$ is the weight output by source s_k; $\epsilon \sim \mathcal{N}(0, \boldsymbol{\Lambda})$ is noise of the system. Parameters ($\boldsymbol{\theta}$) of the model are: $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_K\}$ where π_k is the probability of $s_k = 1$; a $D \times K$ weight matrix $\mathbf{W}$ whose column $\mathbf{w}_k$ corresponds to the weight output for source s_k; $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_K\}$ whose components are the precision ($\frac{1}{\sigma^2}$) of columns of the weight matrix; the covariance matrix $\boldsymbol{\Lambda} = \tau^{-1}\mathbf{I}$ where τ is the precision of noise ϵ. The graphic representation of the model is shown in Figure 1. The learning task includes the parameter estimation and model selection based on the Bayesian framework.

3 Bayesian Model Selection

The main task of model selection in the VBCVQ model is to determine the number of processes (sources) in the model. In the Bayesian model selection framework, we choose the model $\mathcal{M}_i$ with the highest posterior probability $P(\mathcal{M}_i|\mathbf{Y})$ among a set of models, ($\mathcal{M} = \{\mathcal{M}_j\}_{j=1}^{M}$), based on the observed data. Therefore the selection of the model is dictated by observed data, not arbitrarily by the modeler. According to Bayes' theorem, the posterior probability of a model equals:

$$\begin{aligned} P(\mathcal{M}_i|\mathbf{Y}) &= \frac{P(\mathbf{Y}|\mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathbf{Y})} && (1) \\ P(\mathbf{Y}|M_i) &= \int_{\boldsymbol{\theta}} P(\mathbf{Y}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i)d\boldsymbol{\theta} && (2) \end{aligned}$$

where $\mathbf{Y} = \{\mathbf{y}^{(n)}\}_{n=1}^{N}$ are the observed data; $P(\mathbf{Y}|\mathcal{M}_i)$ is the marginal likelihood or "evidence" for the model; $P(\mathcal{M}_i)$ is the prior probability for the model $\mathcal{M}_i$. If no prior knowledge is available, we use an uninformative prior $P(\mathcal{M}_i)$ and the model selection is determined by $P(\mathbf{Y}|\mathcal{M}_i)$.

Variational approximations. The evaluation of equation (2) is often intractable in practice. Various techniques are used to approximate the integration, e.g., Laplace approximation, Bayesian information criteria (BIC) and Markov Chain Monte Carlo (MCMC) simulation [13]. Recently, the variational Bayesian approach has been used in various statistical models to approximate the integration in equation (2) [15,16,12,10]. The approach takes advantage of the fact that, for a given model M_i, the log marginal likelihood, $\ln P(\mathbf{Y}|M_i)$, can be bounded from below [15,12] as:

$$\begin{aligned} \ln P(\mathbf{Y}|M_i) &= \ln \int_{\theta} \sum_{\mathbf{H}} P(\mathbf{Y}, \mathbf{H}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i)d\boldsymbol{\theta} && (3) \\ &\geq \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q(\mathbf{H}, \boldsymbol{\theta}) \ln \frac{P(\mathbf{Y}, \mathbf{H}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i)}{Q(\mathbf{H}, \boldsymbol{\theta})} d\boldsymbol{\theta} \equiv \mathcal{F} && (4) \end{aligned}$$

where $Q(.)$ is an arbitrary distribution, $\mathbf{H}$ and $\boldsymbol{\theta}$ denote sets of hidden variables and parameters of a given model respectively. The inequality is established by Jensen's inequality. Thus, one can treat the lower bound $\mathcal{F}$ as the function of the free distribution $Q(\mathbf{H}, \boldsymbol{\theta})$ and maximize $\mathcal{F}$ with respect to $Q(\boldsymbol{H}, \boldsymbol{\theta})$. The best result is achieved if $Q(\mathbf{H}, \boldsymbol{\theta})$ equals the posterior joint distribution over hidden variables H and parameters θ. However, the evaluation of the true posterior distribution is intractable in most practical cases. To overcome the difficulty, a variational approximation can be achieved by restricting the maximization $Q(\mathbf{H}, \boldsymbol{\theta})$ to a smaller family of distributions chosen for convenience. A common approach is to use the mean-field approximation, which maximized on the family of models in which hidden variables and parameters are independent. Then the joint distribution can be fully factored: $Q(\mathbf{H},\boldsymbol{\theta}) = \prod_{i=1}^{K} Q_H(H_i)\prod_{j=1}^{P} Q_\theta(\theta_j)$. Restricting $Q(H,\theta)$ to this family gives a less tight bound in equation (4), but one can analytically maximize the lower bound of the log marginal likelihood with respect to the factorized family of distributions by an iterative algorithm similar to the EM algorithm[12].

In the Bayesian framework, the parameters of a given model are treated as random quantities, requiring us to specify prior distributions $P(\boldsymbol{\theta}|\mathcal{M}_i)$ for all model parameters. We choose the following conjugate priors to facilitate the estimation of approximate posterior distributions:

$$P(\boldsymbol{\pi}) = \prod_{k=1}^{K} Beta(\pi_k|\alpha,\beta); \qquad P(\mathbf{W}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{w}_k|0,\gamma_k);$$

$$P(\boldsymbol{\gamma}) = \prod_{k=1}^{K} \mathcal{G}(\gamma_k|a_\gamma,b_\gamma); \qquad P(\tau) = \mathcal{G}(\tau|c_\tau,d_\tau);$$

where $Beta(.|\alpha,\beta)$ is a beta distribution; $\mathcal{G}(.|a,b)$ is a gamma distribution. We use the following set of values of hyper-parameters: $\alpha=\beta=1$, $a_\gamma=b_\gamma=c_\tau=d_\tau=10^{-3}$ during training sessions.

4 Variational Bayesian Learning

In the variational Bayesian approach, we maximize the lower bound $\mathcal{F}$ of the marginal log likelihood $\ln P(\mathbf{Y}|\mathcal{M}_i)$ with respect to a set of parameterized variational distributions $Q(H_k), k=1,2,\ldots,K$ and $Q(\theta_p), p=1,2,\ldots,P$, which are approximate posterior distributions of hidden variables and parameters[15,12]. The process of maximizing the lower bound $\mathcal{F}$ and learning parameter is very similar to conventional expectation-maximization (EM) algorithm[11]. We adopt iterative variational approximation principle[15,12], which maximizes the function $\mathcal{F}$ by iterating over two alternating re-estimation steps:

- Estimation of hidden source distributions $Q_H(\mathbf{H})$:

$$Q_H^*(\mathbf{H}) \propto \exp\langle \ln P(\mathbf{Y},\mathbf{H}|\boldsymbol{\theta})\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{5}$$

- Estimation of parameter posteriors $Q_\theta(\boldsymbol{\theta})$

$$Q_\theta^*(\boldsymbol{\theta}) \propto P(\boldsymbol{\theta})\exp\langle \ln P(\mathbf{Y},\mathbf{H}|\boldsymbol{\theta})\rangle_{Q_H(\mathbf{H})} \tag{6}$$

where $\langle . \rangle_{Q(.)}$ denotes the expectation w.r.t. distribution $Q(.)$.

Expanding and evaluating the equations (5) and (6), we obtain a set of approximate posterior distributions of the hidden sources $\mathbf{H}$ and parameters $\boldsymbol{\theta}$. Thus, the variational Bayesian approach allows us not only to approximate the log marginal likelihood $\ln P(\mathbf{Y}|\mathcal{M}_i)$ to achieve model selection, but also to learn the approximate distributions of the parameters. In the following, we summarize the form of the approximate posterior distributions and rules of updating the parameters of the distributions. Complete derivations can be found in the separate report[17].

$$Q(\mathbf{s}) = \prod_{k=1}^{K} Be(s_k|\lambda_k); \qquad Q(\boldsymbol{\pi}) = \textstyle\prod_{k=1}^{K} Beta(\pi_k|\tilde{\alpha}_k, \tilde{\beta}_k);$$

$$Q(\mathbf{W}) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{w}_d|\tilde{\mathbf{m}}_w^{(d)}, \widetilde{\mathbf{\Sigma}}_w^{(d)}\right); \qquad Q(\boldsymbol{\gamma}) = \textstyle\prod_{k=1}^{K} \mathcal{G}(\gamma_k|\tilde{a}_{\gamma k}, \tilde{b}_{\gamma k});$$

$$Q(\tau) = \mathcal{G}(\tau|\tilde{c}_\tau, \tilde{d}_\tau);$$

where $Be(.|\lambda)$ is a Bernoulli distribution. One can maximize the lower bound $\mathcal{F}$ by initializing the parameters of the model with a suitable guess, then iteratively update the parameters for individual approximate distribution using following updating rules until $\mathcal{F}$ converges to a local maximum.

$$\begin{aligned} \ln \frac{\lambda_k}{(1-\lambda_k)} &= \left\langle \ln \frac{\pi_k}{(1-\pi_k)} \right\rangle + \mathbf{y}^T \left\langle \mathbf{\Lambda}^{-1} \right\rangle \langle \mathbf{w}_k \rangle - \sum_{j \neq k} \lambda_j tr\left(\left\langle \mathbf{w}_j \mathbf{w}_k^T \right\rangle \left\langle \mathbf{\Lambda}^{-1} \right\rangle\right) \\ & -\frac{1}{2} tr\left(\left\langle \mathbf{w}_k \mathbf{w}_k^T \right\rangle \left\langle \mathbf{\Lambda}^{-1} \right\rangle\right); \end{aligned}$$

$$\tilde{\alpha}_k = \alpha_k + \sum_{n=1}^{N} \left\langle s_k^{(n)} \right\rangle; \qquad \tilde{\beta}_k = \beta_k + N - \textstyle\sum_{n=1}^{N} \left\langle s_k^{(n)} \right\rangle;$$

$$\widetilde{\mathbf{\Sigma}}_w^{(d)} = \left(diag(\langle \boldsymbol{\gamma} \rangle) + \langle \tau \rangle \sum_{n=1}^{N} \left\langle \mathbf{s}^{(n)} \mathbf{s}^{(n)T} \right\rangle \right)^{-1}; \qquad \tilde{\mathbf{m}}_w^{(d)} = \widetilde{\mathbf{\Sigma}}_w^{(d)} \langle \tau \rangle \textstyle\sum_{n=1}^{N} \left\langle \mathbf{s}^{(n)} \right\rangle y_d^{(n)};$$

$$\tilde{a}_{\gamma k} = a_{\gamma k} + \frac{D}{2}; \qquad \tilde{b}_{\gamma k} = b_{\gamma k} + \tfrac{\left\langle ||\mathbf{w}_k||^2 \right\rangle}{2};$$

$$\tilde{c}_\tau = c_\tau + \frac{ND}{2};$$

$$\tilde{d}_\tau = d_\tau + \tfrac{1}{2} \textstyle\sum_{n=1}^{N} \left\{ ||\mathbf{y}^{(n)}||^2 - 2\mathbf{y}^{(n)T} \langle \mathbf{W} \rangle \left\langle \mathbf{s}^{(n)} \right\rangle + tr\left(\left\langle \mathbf{W}^T \mathbf{W} \right\rangle \left\langle \mathbf{s}^{(n)} \mathbf{s}^{(n)T} \right\rangle\right) \right\};$$

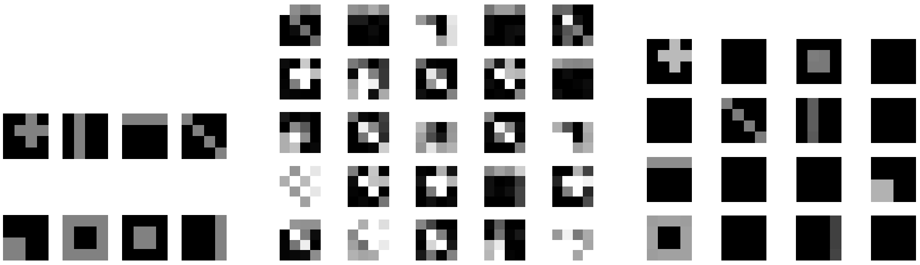

Figure 2: **Left panel**: Original source images used to generate data. **Middle panel**: Observed images resulting from mixture of sources. **Right panel**: Recovered sources

5 Analysis of Simulated Data

We have implemented the variational Bayesian inference algorithm for the CVQ model. To demonstrate the capability of the model to identify the source processes uniquely, we first applied the model to a simulated microarray data.

In this experiment, we used 8 hidden sources to simulate cellular processes that control expression of 16 genes. The left panel of Figure 2 depict the components of the model, where genes are represented by pixels of a 4×4 image. Each of the 8 sources controls a subset of 16 genes, where the intensity of the pixels reflect the degree of influence by the source. As the figure shows, some genes are controlled by multiple sources. We generated 600 images (experimental data) by setting sources to be "on/off" stochastically, summing the weight output by sources and adding random noise into the images. The middle panel of Figure 2 illustrates some of the data images generated during the process. We run our program to test its ability of automatically recovering the number of sources and their patterns. The right panel of Figure 2 shows the result of an experiment where the algorithm is initialized with 16 hidden sources. The program correctly identified all 8 sources that were used to generate the data and eliminated the rest 8 unnecessary sources. The experiment demonstrates an excellent performance of the variational Bayesian approach on blind source separation for simulated gene expression data.

Figure 2 also shows an interesting characteristic of our Bayesian CVQ model – its ability to eliminate unnecessary sources automatically, thus, achieving the effect of model selection. Such an ability is due to the introduction of hierarchical parameters γ (see Section 2) into the model. The approach is referred to as automatic relevance determination (ARD). It has been used in a number of Bayesian linear latent variable models to determine model dimension automatically. [16,18,10].

When variational Bayesian ICA model with mixture of Gaussian sources was first tested to perform a similar image separation task[19,10], recovery of source images

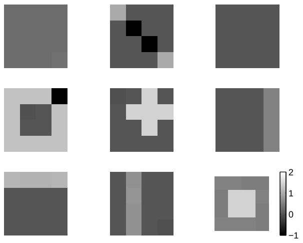

Figure 3: Source processes recovered from the training data containing a background signal and both positive and negative weight sources. The first image captures the background signal. Black pixels capture negative weights.

from the mixed image data was hindered by contamination with negative "ghost" images. In order to prevent "ghost" images, special constraints on distributions were incorporated into the ICA model. Specifically, the use of rectified Gaussian distributions priors[10] restricted both the source and weight matrix to the positive domain. In contrast, the CVQ model performs blind source separation without special constraints. Adopting Bernoulli distributions for sources in the CVQ model naturally constrains the sources to the non-negative domain, preventing "ghost" images. No constraint on the weight matrix appears necessary. This flexibility allows the capture of *genuine* negative influences of sources on the observed data, which is a highly desirable characteristic for detecting the repressive effects of signal transduction components on gene expression. To test the model's ability to capture repressive effects, we generated 600 training data with 8 sources similar to those described earlier with one exception: weight outputs for two sources are negative on some of the pixels. We randomly initialized parameters for hidden sources, and then ran the algorithm to recover the sources. Once again our variational Bayesian algorithm was able to identify correctly not only the number of underlying regulatory signals but also their weight matrices, including their repressive (negative) components. Figure 3 shows the sources and weights recovered by the algorithm for the simulated data. Black pixels correspond to negative weights.

6 Application in Microarray Data Analysis

In this section, we present the result of applying the CVQ data analysis to the yeast cell cycle data by Spellman et al[20]. These cell cycle data has been widely used to test different algorithms, including SVD and ICA[1,3]. The data set contains a collection of the whole yeast genome expression measurements (77 samples) across the yeast cell cycle. During the cell cycle, the states of the cellular processes that controls progression of cell cycle switch “on/off” periodically. Thus, these data are suitable to test the ability of the CVQ model to capture such periodical behavior of cellular processes.

We have extracted expression patterns of 697 genes that are documented to be cell-cycle dependent [20] and used the CVQ to model the data. Original data is in the form of log ratio of fluorescence of labeled sample cDNA and control cDNA. Before fitting the model, the log ratio of the data was transformed to positive values by subtract the minimum ratio of each gene. In order to determine the optimal model that fit the data well, we tested CVQ models setting the initial number of sources to values ranging from 8 to 30. We ran each model 30 times. Figure 4 shows the results of experiments. We can see that the lower bound $\mathcal{F}$ for log marginal likelihood reaches a plateau between the models with 12 to 20 sources. Inspecting the recovered models, we found that most of these models have 12 working sources; excess sources were eliminated by the ARD phenomenon. Note that models initialized with more than 20 sources are penalized by the Bayesian approach in that the $\mathcal{F}$ values begin to drop. Thus, the variational Bayesian approach consistently returned models with 12 sources as the most suitable model for the observed data. In comparison to the models studied by Alter et al [1] and Liebermeister [3], where the number of processes was determined by the number of samples, our approach determines the number of processes based on the sound statistical foundation of the Bayesian framework. In addition, the larger number of processes in their model significantly increases the number of parameters to estimate – about 50,000 more parameters would be needed to carry out a similar experiment. It is well known that the models with a large number of parameters are prone to over-fitting the training data, especially with a training set of a small size like the one used in our experiment. The full Bayesian treatment of the CVQ model implicitly penalizes models with too many parameters, thus making it less likely to over-fit the data.

We have studied the recovered CVQ model to see if it can capture the periodic behaviors of the processes. The middle and right panel of the Figure 4 show one of the recovered models with the highest $\mathcal{F}$. The middle panel shows the state of 12 hidden sources across the experiment conditions, in this case, a times series of gene expression observations. One can clearly see the cyclic “on/off” pattern of the sources which are far from being random. This is not surprising and encouraging, as we are modeling expression control processes of cell cycle related genes. For each of the cell cycle time points, we can see sources cooperatively contributing to observations. Thus, the CVQ model provides another approach to decomposing the overall observation at genome level into different processes, which may reflect the state of different cellular signal transduction components. A more detailed biological analysis of the results is being carried out and will be reported separately.

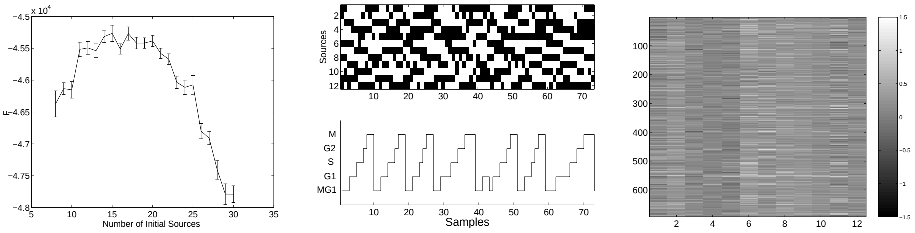

Figure 4: **Left panel**: Mean and standard deviation of $\mathcal{F}$ of models initialized with different number of sources. **Middle panel**:*Top*: States of hidden sources (rows) of each time series observations (columns). Black blocks indicate the source is "on" and white blocks indicate the source is "off". *Bottom*: Corresponding cell cycle phase for each observation. **Right panel**: The weights associated with sources (columns).

7 Discussion

One important aspect of systems biology is to understand how information is organized inside the cell. For example, an interesting question is: what is the minimum number of central signal transduction components needed to coordinate the variety of cellular signals and cellular function. A cell is constantly bombarded by extracellular signals; many of these signals are eventually propagated to the nucleus in order to regulate gene expression. It would be surprisingly inefficient for nature to endow every receptor at the plasma membrane with a unique pathway to pass its signal from plasma membrane to the promoter of a gene. Rather more plausible is a minimum set of partially shared signal transduction components that play central role in coordinating signals from extracellular environment and disseminating the signals to the transcription factor level. These components work as encoders that compress a large amount of information from extracellular and intracellular environments to minimum length, then pass the information to gene expression regulating components such as transcription factors or repressors. To model these signal transduction components, model selection becomes a key issue, which has not been well addressed previously. Bayesian model selection respects Occam's razor, to minimize a fitted model's complexity, potentially increase the interpretability of the data in terms of information organization and flow inside living cells. These characteristics put the model a step ahead of some commonly used models for modeling cellular processes controlling gene expression.

Like most other models used to decompose observed microarray data into components, the CVQ model is a linear model. In microarray data analysis, measurements are usually transformed by the logarithm, so that cooperative effects that combine multiplicatively at the raw data level can be handled as additive. This simplifies model-fitting but may be too restrictive. To capture nonlinear relationships in the log space, the CVQ model could naturally be extended to mixtures of CVQ models. This extension will be studied in the future. Another possible improvement of the model includes more sophisticated approximation methods, such as Minka's expectation propagation method [21], to obtain a better approximation of the log marginal likelihood, and thus, better model selection and optimization.

Acknowledgments

The authors would like to extend special thanks to Dr. Zoubin Ghahramani for constructive initial inputs and discussion. We thank Drs. Gregory Cooper, Chengxiang Zhai, Rong Jin, Vanathi Gopalakrishnan, Matthew J. Beal and two anonymous reviewers for insightful discussions and comments. Xinghua Lu would like to thank for the support from the National Library of Medicine (training grant 3T15 LM07059-15S1) and the MUSC COBRE for Cardiovascular Disease.

Reference

1. Alter, O, Brown, P. O. and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of Ameerica*, 97:10101–10106, 2000.
2. Raychaudhuri, S., Stuart, J. M. and Altman, R. B.. Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Proceeding of Pacific Symposium on Biocomputing*, pages 455–66, 2000.
3. Liebermeister, W. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18:51–60, 2002.
4. Martoglio, A, Miskin, J. W., Smith, S. K. and MacKay, D. J. C.. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*, 18 no. 12:1617–1624, 2002.
5. Moloshok, T. D., Klevecz, R. R., Grant, J. D., Manion, F. J., Speier W. F., and Ochs, M. F.. Application of Bayesian decomposition for analysing microarray data. *Bioinformatics*, 18(4):566–575, 2002.
6. Segal, E, Battle, A and Koller, D. Decomposing gene expression into cellular processes. In *Proceedings of Pacific Symposium on Biocomputing*, volume 8, pages 89–100, 2003.
7. Attias, H.. Independent Factor Analysis. *Neural Computation*, 11(4):803–851, 1999.
8. Hinton, G. E. and Zemel, R. S.. Autoencoders, minimum description length, and helmholtz free energy. In *Advances in Neural Information Processing Systems 6*. Morgan Kaufman, 1994.
9. Ghahramani, Z. Factorial learning and EM algorithm. In *Advances in Neural Information Processing Systems 7*. Morgan Kaufmann Publishers, 1995.
10. Miskin, J and MacKay, D. Ensemble learning for blind source separation. In S. Roberts and R. Everson, editors, *Independent Component Analysis: Principles and Practice*, pages 209–233. Cambridge Unviersity Press, 2001.

11. Dempster, A.P., Laird, N.M. and Rubin, D.B.. Maximum likelihood estimation from incomplete data via EM algorithm (with discussion). *Journal of Royal Statistics Society*, B 39:1 – 38, 1977.
12. Ghahramani, Z and Beal, M. J.. Propagation algorithms for variational bayesian learning. In *Advances in Neural Information Processing Systems 12,* pages 507–513. MIT Press, 2000.
13. Kass, R and Raftery, A, E.. Bayes Factors. Technical Report Technical Report No 254, Dept. of Statistics and Techical Report No 571, Dept. of Statistics, Univ. of Washington and Carnegie Mellon Univ., 1994.
14. MacKay, D. Probable networds and plausible predictions - a review of practical Baysian methods for supervised nerual networkds. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.
15. Attias, H. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Uncertainty in AI Conference*, pages 21–30, 1999.
16. Bishop,C. M.. Variational principal components. In *Proceedings of Ninth International Conference on Artificial Neural Networks*, volume 1, pages 509–514. ICANN, 1999.
17. Lu, X., Hauskrecht, M. and R. S. Day, R. S.. Variational Bayesian learning of cooperative vector quantizer model – theory. Technical Report No: CBMI-02-181, The Center for Biomedical Informatics, University of Pittsburg, 2002.
18. Ghahramani, Z. and Beal, M. J.. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA, 2000. MIT Press.
19. Lawrence, N. D. and Bishop, C. M.. Variational Bayesian independent component analysis. Technical report, Computer Laboratory, University of Cambridge, 2000.
20. Spellman, P. T., Sherlock, G, Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O. Botstein, D and Futcher, B.. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.
21. Minka, M. P.. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.