Bayesian Sparse Correlated Factor Analysis

Abstract

In this paper, we propose a new sparse correlated factor model under a Bayesian framework that intended to model transcription factor regulation in a cell. Unlike the convention factors model, the factors are assumed to be non-negative and correlated. The correlation is due the the prior knowledge on the structure of the factors. To model the factors, a rectified function and the Dirichlet process mixture (DPM) prior are introduced. Moreover, the factors are . The loading matrix is sparse and since the prior knowledge of non-zero elements are assumed available, the sparse pattern of the loading matrix is significantly constrained, resulting unambiguous factor order. A Gibbs sampler is proposed to uncover the unknown non-negative factors and the loading matrix from data. The model and the Gibbs sampler are validated on the simulated systems.

Index Terms

Bayesian factor analysis (BFA), Gibbs sampling, Dirichlet process mixture (DPM), sparse, correlated

I. Introduction

In many signal processing applications, the observations are resulted by a linear combination of a set of latent variables, or factors [1], [2]. In speech processing, factors are directly related to multiple sound sources or speakers. In EEG application, factors are event-related brain potentials indicative of the activity of brain. Recognizing and understanding the activities of factors are important tasks therein, which are tackled by factor analysis (FA) methods including, most notably, PCA and ICA.

In this paper, we proposed a new factor analysis model that is motivated by the study of reconstructing a transcriptional regulatory networks (TRN) using microarray data. TRN represents genome-wide regulation of mRNA gene expressions by important regulatory proteins known as transcription factors (TFs). Following a common analogy, TFs function as switches that directly turn on and off gene product at mRNA level, resulting also indirect effects on the translational process for the protein synthesis. Understanding and reconstructing TRNs comprised one of the most active research area in genomics signal processing. To reconstruct TRNs in a cell, microarray measurements of gene mRNA expression profile is often used, which represents the output of a TRN to be reconstructed. If assuming gene expression is due to linear combination of TF regulations, a FA model would be a natural choice of modeling with factors representing the unobserved protein-level activities of all the TFs, since simultaneous protein level measurements of TF are not commonly available. The loading matrix in this FA model indicates the strength and the type (up- or down- regulation) of regulation.

However, due to distinct features of TRNs, conventional FA model is not readily applicable. First, since many TFs can either share the same protein complex or regulate each other, the factors should be correlated. However, in the exiting FA models, factors are overwhelmingly assumed independent, which, although true in many applications, is not a realistic assumption for TRNs. Secondly, since a TF regulates a small subset of genes, the loading matrix should be sparse. While with the knowledge of TF regulated genes becomes more complete and increasingly available, the prior probabilities of non-zero elements of the loading matrix is available and should included in modeling. The inclusion of prior for sparsity naturally calls for a Bayesian solution. As an added advantage, having this prior knowledge actually resolves the factor order ambiguity of the conventional factor analysis. Thirdly, the TF activities could be nonGaussian and a nonGaussian factor model should be in place.

In a response to meet these requirements of TRNs, we proposed here a novel Bayesian sparse correlated factor model. The sparsity of loading matrix is constrained by a sparse prior that directly reflects our

knowledge of TF regulation [5]. To model the correlated factors, a Dirichlet process mixture (DPM) prior [4] was placed on the factors. DPM imposes a natural clustering effect on TFs, which, as will be shown, also leads to non-Gaussian distributions on the factors. Moreover, its enables automatic determination of the optimal number of clusters. To effectively estimate the sparse loading matrix and the factor clusters, a Gibbs sampling solution is proposed. Simulation results demonstrate the validity of the model and effectiveness of the proposed Gibbs sampling algorithm.

II. Bayesian Sparse Correlated Factor Analysis

Let $\mathbf{y}_n \in \mathcal{R}^{\mathcal{G}\times\infty}$ for $n = 1, \ldots, N$ represent the nth vector observations, which for TRNs denote the microarray data of G genes. Assume $\mathbf{y}_n$ is generated by the following factor model

$$\mathbf{y}_n = \mathbf{A}\mathbf{x}_n + \mathbf{e}_n \tag{1}$$

$$\mathbf{x}_n = h(\mathbf{s}_n) \tag{2}$$

where

$h(\cdot)$- is the component-wise rectification (or cut) function defined as $h(\cdot) = \max(\cdot, 0)$. The rectification function ensures non-negativity of factors.

$\mathbf{s}_n$- is the vector of pseudo factors, through which the nonnegative factor is defined. $\mathbf{s}_n$ is also assumed to exhibit clustering effect, which is modeled *a priori* by a Dirichlet process mixture (DPM) of Gaussian as

$$s_{l,n} \sim \mathcal{N}(\boldsymbol{\theta}_{l,n}); \quad \boldsymbol{\theta}_{l,n} \sim G; \quad G \sim DP(\alpha, NIG(\boldsymbol{\lambda}_0))$$

where $\boldsymbol{\theta} = \{\mu_{l,n}, \sigma^2_{l,n}\}$, DP denotes the Dirichlet process, and NIG is short for the conjugate Normal-Inverse-Gamma (NIG) distribution with the parameters $\boldsymbol{\lambda}_0 = \mu_0, \kappa_0, \alpha_0, \beta_0$. This DPM prior can also be expressed as

$$p(s_{l,n}|\gamma_l, \boldsymbol{\theta}_{\gamma_l,n}) = \mathcal{N}(\boldsymbol{\theta}_{\gamma_l,n});$$

$$p(\boldsymbol{\theta}_{\gamma_l,n}) \sim NIG(\lambda_0); \quad \gamma_l \sim \boldsymbol{\pi}; \quad \boldsymbol{\pi} \sim GEM(\alpha)$$

where γ_l represents the cluster label of the l-th factor, which is governed by a discrete distribution $\boldsymbol{\pi}$ that follows the GEM distribution [4] that defines the stick breaking process of parameter α. Based on this definition, the conditional distribution

$$\begin{aligned} & p(s_{l,n}, \gamma_l | \mathbf{s}_{-l,n}, \gamma_{-l}) \\ = \quad & (\sum_{k=1}^{K} N_{-l,k} p(s_{l,n} | s_{i,n} \forall i \in \mathcal{S}_{-l,k}, \gamma_l) \delta(\gamma_l - k) \\ & + \alpha p(s_{1,n}) \delta(\gamma_l - \bar{k})) / (\alpha + L - 1) \end{aligned} \tag{3}$$

where, $\gamma_l = k$; $\mathcal{S}_{-l,k} = \{i | i \neq l, \gamma_i = k\}$ represents the set of the factors besides s_l that also belong to cluster k, $N_{-l,k}$ is size of $\mathcal{S}$, and $\bar{k}$ is a new cluster label other than the K existing ones. The distribution (14) demonstrates the correlation between factors - s_l depends only on other factors belonging to the same cluster. The predictive density $p(s_{l,n}|\mathbf{s}_{-l,n}, \gamma_l)$ is shown to be a Student-t distribution, which can be conveniently approximately by the normal distribution for large N_k as

$$p(s_{l,n}|\mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l) \approx \mathcal{N}(\hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n}) \tag{4}$$

where,

$$\gamma_j \in \{1, 2, ..., K, \bar{k}\}$$

$$\hat{\mu}_{l,n} = (\mu_0 \kappa_0 + \sum_{i \in \mathcal{S}_{-l,k}} s_{i,n}) / \bar{\kappa}; \bar{\kappa} = \kappa_0 + N_{-l,k}$$

$$\hat{\sigma}^2_{l,n} = (\bar{\kappa} + 1) \bar{\beta} / \bar{\kappa} / (\alpha_0 + N_{-l,k}/2 - 1)$$

$$\bar{\beta} = \beta_0 + (\sum_{i \in \mathcal{S}_{-l,k}} s^2_{i,n} + \kappa_0 \mu_0^2 - \bar{\kappa} \hat{\mu}^2_{l,n})/2.$$

For the special case, when $\gamma_l = \bar{k}$, the factor prior $p(s_{l,n})$ can be also shown to be a Student-t distribution and approximated by

$$p(s_{l,n}) \approx \mathcal{N}(\mu_0, \sigma^2_{s0}); \sigma^2_{s0} = (\kappa_0 + 1)\beta_0 / \kappa_0 / (\alpha_0 - 1).$$

Although the Gaussian approximation holds well for large α_0, the fact that the factor prior is a Student-t distribution suggests that the factors are nonGaussian in our model. To overcome the scale ambiguity of the factor model, we assume a fixed variance σ_{s0}^2.

$\mathbf{x}_n$- the $L \times 1$ vector of the non-negative correlated factors. For TRNs, $\mathbf{x}_n$ represents the non-negative activity of L TFs of interest. Given the latent factors $\mathbf{s}_n$, the conditional prior of $x_{l,n}$ can be expressed as

$$\begin{aligned} p(x_{l,n}|s_{l,n}) &= \delta(x_{l,n} - h(s_{l,n})) \\ &= U(-s_{l,n})\delta(x_{l,n}) + \delta(x_{l,n} - s_{l,n})U(x_{l,n}) \end{aligned} \tag{5}$$

where $U(\cdot)$ represents the unit step function. This is a more realistic model for TF activities since they should be positive. In the literature, the truncated Gaussian distribution is often adopted to model nonnegative random variables. Compared with the truncated Gaussian distribution, the distribution 5 puts a mass $\tilde{\pi}_{x_{l,n}}$ at $x_{l,n} = 0$. Consequently, it induces sparsity among factors. Moreover, this distribution provides an added convenience that allows analytical solutions for the required conditional distributions in the subsequent Gibbs sampling solution.

$\mathbf{A}$- the $G \times L$ sparse loading matrix. The elements of A are assumed to be independent and with the *a priori* distribution [5]

$$p(a_{gl}) = (1 - \pi_{gj})\delta(a_{gj}) + \pi_{gj}\mathcal{N}(a_{gl}|0, \sigma_{a,0}) \tag{6}$$

where π_{gj} is the *a priori* probability of a_{gl} to be nonzero. For TRNs, databases such as Transfac provide experimentally validated target genes of TFs and this knowledge TF regulation is reflected by setting, for instance, $\pi_{gj} = 0.8$, if TF j is known to regulate gene g, and $\pi_{gj} = 0.8$, otherwise.

$\mathbf{e}_n$ - the $G \times 1$ white Gaussian noise vector with the covariance matrix defined by $\Phi = diag(\sigma_{e,1}^2, \ldots, \sigma_{e,G}^2)$.

The goal of Bayesian analysis is to infer $\mathbf{A}$, $\mathbf{x}_n$, $\mathbf{s}_n$ $\forall n$, $\boldsymbol{\gamma}$ and Φ based on the observations $\mathbf{y}_n$ $\forall n$. For convenience, $\boldsymbol{\Theta}$, $\mathbf{y}_{1:N}$, and $\mathbf{X}$ are introduced to denote the sets of all these unknowns, all the

observations, and all the factors, respectively. Note that the total number of factor clusters K and $\boldsymbol{\theta}_k$ $\forall k$ are also unknown but treated implicitly by the proposed Bayesian solution.

III. The proposed Gibbs sampling solution

Gibbs sampling devises a Markov Chain Monte Carlo scheme to generate random samples of the unknowns from the desired but intractable posterior distributions and then approximate the (marginal) posterior distributions with these samples. The key of Gibbs sampling is to derive the conditional posterior distributions and then draw samples from them iteratively. The intended conditional distributions of the proposed Gibbs sampling solution are detailed next.

1. $\underline{p(a_{gl}|\boldsymbol{\Theta}_{-a_{gl}}, \mathbf{y}_{1,N})}$: Let $\hat{\mathbf{y}}_{gl} = [\hat{y}_{gl,1}, \ldots, \hat{y}_{gl,N}]^\top$ with $\hat{y}_{gl,n} = y_{g,n} - \sum_{i=1, i\neq l}^{L} a_{gi} x_{i,n}$ and $\mathbf{x}_l = [x_{l,1}, \ldots, x_{l,N}]^\top$. It then follows $\hat{\mathbf{y}}_{gl} \sim \mathcal{N}(\hat{\mathbf{y}}_{gl}|\mathbf{x}_l a_{gl}, \sigma_{e,g}^2 \mathbf{I}_N)$ and

$$
\begin{aligned}
& p(a_{gl}|\boldsymbol{\Theta}_{-a_{gl}, \mathbf{y}_{1,N}}) \\
&= p(a_{gl}|\mathbf{x}_l, \hat{\mathbf{y}}_{gl}, \sigma_{e,g}^2) && (7)\\
&= Z_0 p(\hat{\mathbf{y}}_{gl}|\mathbf{x}_l, a_{gl}, \sigma_{e,g}^2) p(a_{gl}) && (8)\\
&= (1-\hat{\pi}_{gl})\delta(a_{gl}) + \hat{\pi}_{gl}\mathcal{N}(a_{gl}|\hat{\mu}_{a,gl}, \hat{\sigma}_{a,gl}^2) && (9)
\end{aligned}
$$

where Z_0 is a normalizing constant, and $\hat{\pi}_{gl} = \pi_{gl}/((1-\pi_{gl})BF_{01} + \pi_{gl})$ is the posterior probability of $a_{gl} \neq 0$ and BF_{01} is the Bayes factor of model $a_{gl} = 0$ vs. model $a_{gl} \neq 0$ and expressed by

$$
\begin{aligned}
BF_{01} &= \frac{p(\hat{\mathbf{y}}_{gl}|\mathbf{x}_l, a_{gl} = 0, \sigma_{e,g}^2))}{p(\hat{\mathbf{y}}_{gl}|a_{gl} \neq 0)} && (10)\\
&= \frac{\mathcal{N}(\hat{\mathbf{y}}_{gl}|\mathbf{0}, \sigma_{e,g}^2 \mathbf{I}_N)}{\mathcal{N}(\hat{\mathbf{y}}_{gl}|\mathbf{0}, \sigma_{e,g}^2 \mathbf{I}_N + \mathbf{x}_l \mathbf{x}_l^\top \sigma_{a,0})} && (11)
\end{aligned}
$$

and $\hat{\mu}_{a,gl} = \hat{\sigma}_{a,gl}^2 \mathbf{x}_l^\top \hat{\mathbf{y}}_{gl}/\sigma_{e,g}^2$ and $(\hat{\sigma}_{a,gl}^2)^{-1} = (\sigma_{a,0})^{-1} + \mathbf{x}_l^\top \mathbf{x}_l/\sigma_{e,g}^2$.

Sampling a_{gl} from (7) resulting a draw of either 0 with probability $1 - \hat{\pi}_{gl}$ or from $f(a_{gl})$ with $\hat{\pi}_{gl}$. Notice that $\hat{\pi}_{gl}$ is inversely associated by $BF_{0,1}$ and when $\pi_{gl} = 0.5$, i.e, a noninformative prior on

sparsity is assumed, $\hat{\pi}_{gl}$ depends only on $BF_{0,1}$ and when $BF_{0,1} > 1$, $\hat{\pi}_{gl} < 0.5$. Since model selection based $BF_{0,1}$ favors $a_{gl} = 0$, it suggests that this Bayesian solution favors sparse model even when $\pi_{gl} = 0.5$.

2. $\underline{p(\gamma_l|\mathbf{\Theta}_{-\mathbf{x}_l,\gamma_l}, \mathbf{y}_{1:N})}$ It should be noted that γ_l does not depend on $\mathbf{x}_l$ in the distribution. It is intended that samples of γ_l from this distribution is not affected by the immediate sample of $\mathbf{x}_l$, thus achieving faster convergence of the sample Markov chains. To derive this distribution, first let $\hat{\mathbf{y}}_{l,n} = \mathbf{y}_n - \mathbf{A}\mathbf{x}_n + \mathbf{a}_l x_{l,n}$ with $\mathbf{a}_l$ being the lth column of $\mathbf{A}$ and hence $\hat{\mathbf{y}}_{l,n} \sim \mathcal{N}(\hat{\mathbf{y}}_{l,n}|\mathbf{a}_l x_{l,n}, \Phi)$. Then,

$$\begin{aligned} & p(\gamma_l|\mathbf{\Theta}_{-\mathbf{x}_l,\gamma_l}, \mathbf{y}_{1:N}) \\ &= p(\gamma_l|\gamma_{-l}, \hat{\mathbf{y}}_{l,1:N}) \\ &= \int p(\gamma_l, \mathbf{x}_l|\gamma_{-l}, \hat{\mathbf{y}}_{l,1:N}) d\mathbf{x}_l \\ &= Z_0 \int p(\hat{\mathbf{y}}_{l,1:N}|\mathbf{x}_l) p(x_{l,n}, \gamma_l|\mathbf{x}_{-l,n}, \boldsymbol{\gamma}_{-l}) d\mathbf{x}_l \\ &= \frac{1}{Z_l}(\sum_{k=1}^{K} N_{-l,k} g_{l,k} \delta(\gamma_l - k) + \alpha g_{l,\bar{k}} \delta(\gamma_l - \bar{k})) \end{aligned} \tag{12}$$

where

$Z_l = \sum_{k=1}^{K} N_{-l,k} g_{l,k} + \alpha g_{l,\bar{k}}$,

$$\begin{aligned} g_{l,k} &= \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{y}}_{l,n}|\mathbf{0}, \Phi) Q(\frac{-\hat{\mu}_{l,n}}{\hat{\sigma}_{l,n}}) \\ &\quad + \mathcal{N}(\hat{\mathbf{y}}_{l,n}|\mu_{\hat{\mathbf{y}}_{l,n}}, \Sigma_{\hat{\mathbf{y}}_{l,n}})[1 - Q(\frac{-\hat{\mu}_{l,n}}{\hat{\sigma}_{l,n}}])) \end{aligned}$$

where, $\mu_{\hat{\mathbf{y}}_{l,n}} = \mathbf{a}_l \mu_{x_{l,n}}$ and $\Sigma_{\hat{\mathbf{y}}_{l,n}} = \mathbf{a}_l \mathbf{a}_l^\top \sigma^2_{x_{l,n}}$ and $g_{l,\bar{k}}$ is a special case of $g_{l,k}$

3. $\underline{p(\mathbf{x}_l|\mathbf{\Theta}_{-\mathbf{x}_l}, \mathbf{y}_{1:N})}$ This distribution can be expressed as

$$
\begin{aligned}
& p(\mathbf{x}_l|\mathbf{\Theta}_{-\mathbf{x}_l}, \mathbf{y}_{1:N}) \\
&= p(\mathbf{x}_l|\boldsymbol{\gamma}_{-l}, \mathbf{s}_{-l}, \hat{\mathbf{y}}_{1:N}, \Phi) \\
&= Z_0 \prod_{n=1}^{N} p(\hat{\mathbf{y}}_n|x_{l,n}) p(x_{l,n}|\mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}) \\
&= Z_0 \prod_{n=1}^{N} p(\hat{\mathbf{y}}_n|x_{l,n}) (\sum_{k=1}^{K} p(x_{l,n}|\mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l = k) + p(x_{l,n}|\mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l = \bar{k})) \\
&= Z_0 \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{y}}_{l,n}|\mathbf{a}_l x_{l,n}, \Phi) ((\sum_{k=1}^{K} \\
& \quad p(x_{l,n}|s_{i,n} \forall i \in \mathcal{S}_{-l,k}, \gamma_l) \delta(\gamma_l - k)) + p(x_{l,n}) \delta(\gamma_l - \bar{k})) \\
&= \prod_{n=1}^{N} \hat{\pi} \delta(x_{l,n}) + (1 - \hat{\pi}) \mathcal{N}(x_{l,n}|\mu_{x_{l,n}}, \sigma^2_{x_{l,n}}) U(x_{l,n})
\end{aligned}
$$

where

$$
\hat{\pi} = \frac{\mathcal{N}(\hat{\mathbf{y}}_{l,n}|\mathbf{0}, \Phi) Q(\frac{-\hat{\mu}_{l,n}}{\hat{\sigma}_{l,n}})}{\mathcal{N}(\hat{\mathbf{y}}_{l,n}|\mathbf{0}, \Phi) Q(\frac{-\hat{\mu}_{l,n}}{\hat{\sigma}_{l,n}}) + \mathcal{N}(\hat{\mathbf{y}}_{l,n}|\mu_{\hat{\mathbf{y}}_{l,n}}, \Sigma_{\hat{\mathbf{y}}_{l,n}})[1 - Q(\frac{-\hat{\mu}_{l,n}}{\hat{\sigma}_{l,n}})]}
$$

where

$$
\mu_{\hat{\mathbf{y}}_{l,n}} = \mathbf{a}_l \mu_{x_{l,n}}
$$

$$
\Sigma_{\hat{\mathbf{y}}_{l,n}} = \mathbf{a}_l \mathbf{a}_l^\top \sigma^2_{x_{l,n}}
$$

$$
\mu_{x_{l,n}} = \sigma^2_{x_{l,n}} \mu^*
$$

$$
\sigma^2_{x_{l,n}} = (\mathbf{a}_l^\top \Phi^{-1} \mathbf{a}_l + (\hat{\sigma}^2_{l,n})^{-1})^{-1}
$$

$$
\mu^* = \mathbf{a}_l^\top \Phi^{-1} \hat{y}_{l,n} + (\hat{\sigma}^2_{l,n})^{-1} \hat{\mu}_{l,n}
$$

4. $\underline{p(s_{l,n}|\mathbf{\Theta}_{-\mathbf{x}_l}, \mathbf{y}_{1:N})}$ The conditional probability is:

$$
\begin{aligned}
p(x_{l,n}|s_{l,n}) &= \delta(x_{l,n}) U(-s_{l,n}) + \delta(x_{l,n} - s_{l,n}) U(s_{l,n}) \\
&= \tilde{\pi}_{x_{l,n}} \delta(x_{l,n}) + (1 - \tilde{\pi}_{x_{l,n}}) \delta(x_{l,n} - s_{l,n})
\end{aligned}
$$

where, $\tilde{\pi}_{x_{l,n}} = U(-s_{l,n})$

Since (4), the posterior distribution is:

$$\begin{aligned} p(s_{l,n}|x_{l,n}, \mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l) &= (1 - \bar{\bar{\pi}}_{x_{l,n}})\mathcal{N}(x_{l,n}|\hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n})U(-s_{l,n}) \\ &\quad + \bar{\bar{\pi}}_{x_{l,n}}\delta(s_{l,n} - x_{l,n}) \end{aligned}$$

where

$$\begin{aligned} \bar{\bar{\pi}}_{x_{l,n}} &= \frac{\mathcal{N}(x_{l,n}|\hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n})}{\delta(x_{l,n})Q(\frac{-\hat{\mu}_{l,n}}{\hat{\sigma}_{l,n}}) + \mathcal{N}(x_{l,n}|\hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n})U(x_{l,n})} \\ &= sgn(x_{l,n}) \end{aligned}$$

5. $\underline{p(\sigma^2_{e,g}|\boldsymbol{\Theta}, \mathbf{y}_{1:N})}$ Let $\hat{y}_{g,n} = y_{g,n} - \mathbf{a}_g^\top \mathbf{x}_n$ and thus $p(\hat{y}_{1:N}|\sigma^2_{e,g}) = \mathcal{N}(\mathbf{0}, \sigma^2_{e,g}\mathbf{I}_N)$. Given the conjugate prior,

$$\begin{aligned} p(\sigma^2_{e,g}|\boldsymbol{\Theta}, \mathbf{y}_{1:N}) &= p(\sigma^2_{e,g}|\hat{y}_{g,1:N}) \\ &= Z_0 \prod_{n=1}^{N} p(\hat{y}_{1:N}|\sigma^2_{e,g})p(\sigma^2_{e,g}) \\ &= IG(\alpha_g, \beta_g) \end{aligned} \tag{13}$$

where $\alpha_g = \alpha_0 + N/2$ and $\beta_g = \beta_0 + \sum_{n=1}^{N} \hat{y}^2_{g,n}/2$.

The proposed Gibbs sampler can be summarized as follows.

Gibbs Sampling for BSCFA

Iterate the following steps and for the tth iteration

1) Sample $a^{(t)}_{gl}$ $\forall g, l$ from (7);
2) for l=1 to L

   Sample $\gamma^{(t)}_l$ from (12); Set $K = K + 1$ if $\gamma^{(t)}_l = K + 1$;

   Sample $\mathbf{x}^{(t)}_l$ from (13) given $\gamma^{(t)}_l$;

3) Sample $\sigma^2_{e,g}$ $\forall g$ from (13). item Remove empty clusters of $\mathbf{x}$ and reduce K accordingly.

The convergence is determined based on a scheme described in section 11.6 of [3]. The samples after convergence will be collected to approximate the marginal posterior distributions and the Bayesian estimates of the unknowns.

IV. RESULTS

The proposed approach was evaluated on simulated system. The toy data was constructed to mimic a microarray experiment that measures the expression profiles of 50 genes (at 15 sample times), which are regulated by 15 transcription factors that belong to 2 clusters, each of probability 0.6 and 0.4. We assume the density of non-zero elements in $\mathbf{A}$ is 0.2, and the hyperparameters are: $\mu_0 = 1, \sigma_{x0}^{-2} = 0.25, \kappa_0 = 0.2, \alpha_0 = 101; \beta_0 = \kappa_0/(1+\kappa_0)(\alpha_0 - 1)\sigma_{x0}$.

Experiments are developed to testify the convergence of proposed Gibbs sampling approach, the precision-recall-curve of non-zeros elements in $\mathbf{A}$, and the estimation of x. 10000 samples were computed using Gibbs sampler (Fig. 1), since the convergence is reached after around 5000 iterations, the first half was discarded to allow for the samplers to burn in. The precision and recall curve for the identification of non-zero elements of $\mathbf{A}$ is depicted in Fig. 2 to demonstrate the capability to retrieve the regulation relationships between transcription factors and genes. The proposed actually demonstrate a very strong ability to retrieve TF-gene regulation. As it can be seen from Fig. 2, When $\sigma_e = 0.5$, 90% regulations can be retrieved with precision larger than 0.8, and when $\sigma_e = 0.7$, 90% regulations can be retrieved with precision larger than 0.7. As for the estimation of x, the performance decreases as the noise increases, and when $\sigma_e = 1$, the noise mean squared error (MSE) is equal to 0.25.

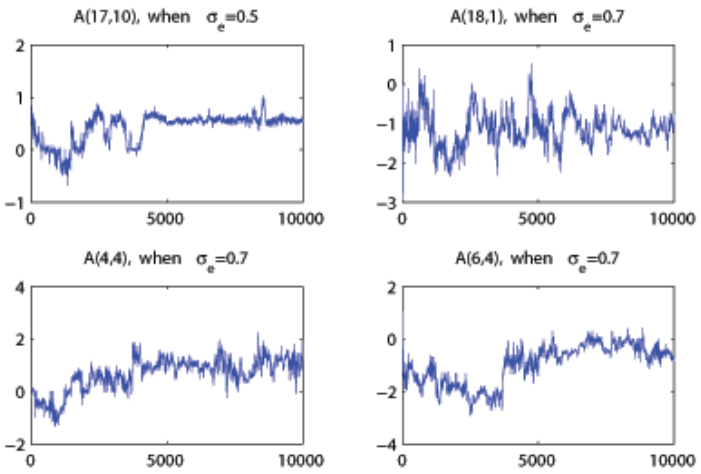

Fig. 1 Gibbs sampling convergence

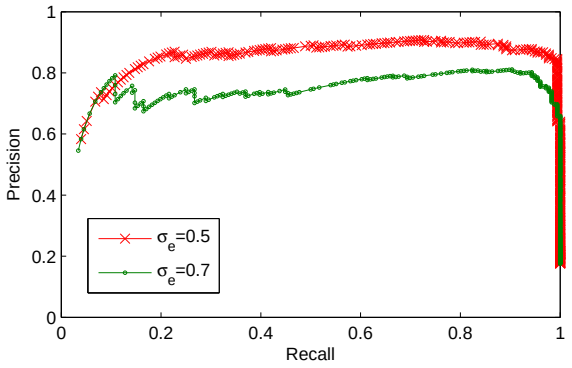

Fig. 2 Identifying Non-Zero Elements of $\mathbf{A}$

V. Conclusion

In this paper, we proposed a new factor model with correlated factors and sparse loading matrix. The factor correlation is modeled by the DPM, which imposes clustering effect on factor. A Gibbs sampler was proposed for resolving the unknown factors and loading matrix. The algorithm has been tested and evaluated on the simulated systems.

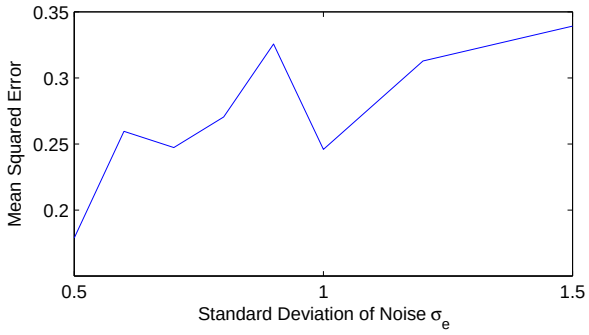

Fig. 3 The Mean Squared Error of the Estimated x

VI. Appendix

A. Derivation of $p(x_{l,n}|\mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l)$

Using the results of (5) and (4), we have

$$\begin{aligned} & p(x_{l,n}|\mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l) \\ = & \int p(x_{l,n}|s_{l,n}) p(s_{l,n}|\mathbf{s}_{-l,n}, \boldsymbol{\gamma}) ds_{l,n} \\ = & Q(-\hat{\mu}_{l,n}/\hat{\sigma}_{l,n})\delta(x_{l,n}) \\ & + \mathcal{N}(x_{l,n}|\hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n}) U(x_{l,n}) \end{aligned} \tag{14}$$

where

$$\hat{\mu}_{l,n} = \frac{\mu_0\kappa_0 + \sum_{i\in S_{-l,k}} s_{i,n}}{\overline{\kappa}}$$

$$\overline{\kappa} = \kappa_0 + N_{-l,k}$$

$$\hat{\sigma}^2_{l,n} = \frac{(\overline{\kappa}+1)\overline{\beta}}{\overline{\kappa}(\alpha_0 + \frac{N_{-l,k}}{2} - 1)}$$

$$\overline{\beta} = \beta_0 + \frac{\sum_{i\in S_{-l,k}} s^2_{i,n} + \kappa_0\mu_0^2 - \overline{\kappa}\hat{\mu}^2_{l,n}}{2}$$

where, $\gamma_l = k$; $S_{-l,k} = \{i | i \neq l, \gamma_i = k\}$ and $N_{-l,k}$ is the size of $S_{-l,k}$.

REFERENCES

[1] J.V. Stone, "A Brief Introduction to Independent Component Analysis in Encyclopedia of Statistics in Behavioral Science," Volume 2, pp. 907C912, Editors Brian S. Everitt & David C. Howell, John Wiley & Sons, Ltd, Chichester, 2005.

[2] Lee," Independent component analysis: Theory and applications," Boston, Mass: Kluwer Academic Publishers, 1998

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*, Chapman and Hall, 2ed, 2008.

[4] Erik B. Sudderth, "Graphical Models for Visual Object Recognition and Tracking," *PhD Thesis*, Massachusetts Institute of Technology.

[5] C. Carvalho, J. Chang, J. Lucas J. R. Nenins, Q. Wang and M. West, "High-Dimensional Sparse Factor Modelling: Applicatons in Gene Expression Genomics," in *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1438-1456, 2008