EUSIPCO 2013 1569743797

BAYESIAN SPARSE FACTOR MODEL FOR TRANSCRIPTIONAL REGULATORY NETWORKS INFERENCE

M. Sanchez-Castillo[†,1] *, I. Tienda-Luna*[3]*, D. Blanco*[1]*, M. C. Carrion-Perez*[1] *and Y. Huang*[3]

[1] Department of Applied Physics, University of Granada, Spain
[3] Department of Electronics, University of Granada, Spain
[3] Department of Electrical and Computer Engineering, University of Texas at San Antonio, USA

ABSTRACT

Uncovering transcription factor (TF) mediated regulatory networks from microarray expression data and prior knowledge is considered in this paper. Bayesian factor models that model direct TF regulation are formulated. To address the enormous computational complexity of the model in large networks, a novel, efficient basis-expansion factor model (BEFaM) has been proposed, where the loading (regulatory) matrix is modeled as an expansion using basis functions of much lower dimension. Great reduction is achieved with BEFaM as the inference involves estimation of expansion coefficients with much reduced dimensions. We also address the issue of incorporating the prior knowledge of TF regulation to constrain the factor loading matrix. A Gibbs sampling solution has been developed to estimate the unknowns. The proposed model was validated by simulation and then applied to breast cancer data to uncover the corresponding TF regulatory network and theirs protein levels.

Index Terms— Bayesian Inference, Gene Expression, Sparse Networks, Transcriptional Networks, Breast Cancer.

1. INTRODUCTION

The development of cells and theirs responses to different stimuli is governed by complex genetic regulatory mechanisms. Gene transcription, the earliest stage of gene regulation, is mediated by a kind of proteins known as transcription factors (TFs) that recognize and bind specific regions of the genes. Uncovering the details of gene regulation and how it defines cellular states and eventually phenotypes is a major challenge facing computational systems biologists. With the accumulation of high throughput genomics data such as microarray expression profiles and biological knowledge, including TF regulated gene sets and gen-protein interaction databases, the development of robust computational models able to fully utilize the data and the prior knowledge comprises one of the active topic in accurate uncovering gene regulations.

[†]Corresponding author: mscastillo@ugr.es

In this paper, we consider the problem of uncovering TF mediated regulatory networks based on gene expression data and prior knowledge of gene regulation. Currently, a large number of models have been proposed including the ordinary differential equations, (probabilistic) Boolean networks, Bayesian networks and information theory based models [1]. Ideally, the TF protein level expression is needed for inferring its context specific regulatory impact. However, due to the low protein coverage and poor quantification accuracy of current proteomics technologies, the measurements of TF protein expressions are hardly available. As a compromise, most of the aforementioned models equate TF mRNA expression to its protein activity. Since gene mRNA expression and its protein expression are far from being correlated, due to post-transcriptional regulation, such treatment is inappropriate and thus the models based on such assumption cannot accurately capture the TF regulatory impact.

In contrast, factor models based on approaches such as network component analysis [2] and Bayesian sparse factor regulatory model [3] treat TF activities as the unknown factors to be estimated and the mRNA expressions as a linear combination of unknown TF activities. These factor models directly address the mechanism of TF-mediated regulations and therefore can result in regulatory networks closer to the reality [4]. Moreover, the uncovered networks by factor models provide information of direct gene regulations by TFs. Despite these appealing features of factor models, they become cumbersome to be applied to data involving a large number of genes and TFs because the number of the unknown parameters for a large system increases exponentially and the inference for such large systems is computationally extremely challenging.

To overcome the aforementioned problems of the conventional factor models, we propose in this paper a novel basis-expansion factor model (BEFaM) where the loading matrix efficiently describes the sparsity properties of the transcriptional regulation and it is modeled as an expansion of basis functions of much lower dimension. The inference for this BEFaM involves the estimation of the expansion coefficients with much reduced dimensions.

We also address the issue of incorporating prior knowledge of the TF regulation to constrain the factor loading matrix. The proposed model was validated by the simulated system and then applied to a real genomic data set of breast cancer to uncover the context specific TF mediated regulatory network. The proposed basis expansion representation of regulatory (loading) matrices significantly reduces model complexity and enables application of factor models to large networks.

2. SPARSE BASIS-EXPANSION FACTOR MODEL

Let $\mathbf{Y} \in \mathbb{R}^{G\times N}$ be a gene expression data set, the log-transformed mRNA gene expression fold-changes versus control, with G genes and N samples. Likewise, consider $\mathbf{X} \in \mathbb{R}^{F\times N}$ the N respective activity profiles of F TFs. We assume that gene expression levels are due a linear combination of the TF activities as

$$\mathbf{y}_n = \mathbf{A}\mathbf{x}_n + \mathbf{e}_n, \forall n = 1, \ldots, N \tag{1}$$

where $\mathbf{y}_n = [y_{1n}, \ldots, y_{Gn}]^\top$ and $\mathbf{x}_n = [x_{1n}, \ldots, x_{Fn}]^\top$ are respectively the gene expression and the unknown TF activities of the n-th sample, $\mathbf{A} \in \mathbb{R}^{G\times F}$ is an unknown loading coefficients matrix and $\mathbf{e}_n$ is experimental noise, distributed as independent white noise by a zero mean Gaussian (Normal) with variance σ_n^2 as

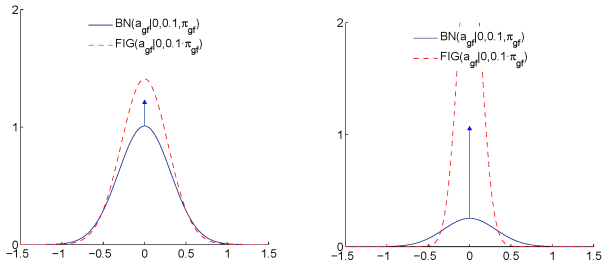
$$p(\mathbf{e}_n) = N\left(\mathbf{e}_n | \, \mathbf{0}^{G\times 1}, \sigma_n^2 \mathbf{1}^G\right). \tag{2}$$

The loading matrix coefficients $\mathbf{a}_f = [a_{1f}, \ldots, a_{Gf}]^\top$ denote the strength with which the f-th TF regulates the expression of each gene. Particularly, these coefficients represent with positive/negative values the up/down regulation or with zero when the TF does not regulate the gene. It is well understood that each TF only regulates a small subset of the genes in a genome and therefore a sparsity representation of the loading matrix should be favored. To this end, Meng et al. have proposed in [3] a sparse Bernoulli Gaussian (BN) priors for the coefficients in $\mathbf{A}$, which impose the probabilities $\pi_{gf}, \forall g, f$ for the coefficients to be non-zero. Therefore, for each coefficient, the sparse modeling is represented by the non-zero prior probabilities, a zero mean and variance σ_f^2 as

$$\begin{aligned} p(a_{gf}) &= BN\left(a_{gf} | \, 0, \sigma_f^2, \pi_{gf}\right) \\ &= (1-\pi_{gf})\,\delta(a_{gf}) + \pi_{gf} N\left(a_{gf} | \, 0, \sigma_f^2\right) \end{aligned} \tag{3}$$

The prior probabilities $\boldsymbol{\pi}_f = [\pi_{1f}, \ldots, \pi_{Gf}]^\top$ represent our prior knowledge of TF f regulating gene g, which can be acquired from, for instance, TransFac database [5]. In this work, to reduce computational complexity, we propose the following Functional prior Induced Gaussian (FIG) distribution to mimic the BN sparse prior as

$$p(\mathbf{a}_f) \approx N\left(\mathbf{a}_f | \, \mathbf{0}, \sigma_f^2 \mathbf{D}_{\boldsymbol{\pi}_f}\right) \tag{4}$$



(a) $\pi_{gf} = 0.8$ (b) $\pi_{gf} = 0.2$

Fig. 1. Bernoulli Gaussian (BN) distribution and its Functional Induced Gaussian (FIG) approximation with zero mean, variance $\sigma_f^2 = 0.1$ and different prior mass probabilities π_{gf}.

where $\mathbf{D}_{\boldsymbol{\pi}_f}$ is a diagonal matrix with elements from vector $\boldsymbol{\pi}_f$. In Figure 1 we represent two BN priors and theirs corresponding FIG approximations for the univariate case. As can be seen, the FIG can closely approximate the BN prior. The FIG has the advantage of being more flexible when modeling high dimensional variables and much more computationally efficient compared with the BN prior. Moreover, it allows the formulation of our proposed BEFaM to be discussed next.

The factor model in (1) includes a large number of unknowns; the TF activities and the (loading) regulatory matrix. The loading matrix $\mathbf{A}$ is the key factor for computational consideration since its dimension increases with $G \times F$ and, for large genome, the size can be in millions. Thus, even with the sparse prior constraint, the number of the unknowns is still extremely large. Inference at such scale is computationally costly expensive and not robust. To address this problem, we propose to model the sparse regulatory matrices with basis expansion as

$$\mathbf{A} = \mathbf{B}\mathbf{C} \tag{5}$$

where $\mathbf{B} \in \mathbb{R}^{G\times K}$ is a matrix of K known bases and $\mathbf{C} \in \mathbb{R}^{K\times F}$ is the coefficient matrix. Note that now we infer $\mathbf{C}$ instead of $\mathbf{A}$ and since $G >> K$, there will be roughly G/K fold reduction in computational complexity with this propose basis expansion. Consequently, the factor model (1) can be expressed as

$$\mathbf{y}_n = \mathbf{B}\mathbf{C}\mathbf{x}_n + \mathbf{e}_n. \tag{6}$$

The efficiency of coefficients $\mathbf{C}$ for modeling the expression data by the proposed expansion model depends on the proposed expansion basis $\mathbf{B}$. According to wavelet theory, any signal can be decomposed into components spanned by the scaling and shifting wavelet basis functions at different resolutions. We consider in this paper the Haar wavelet matrix. Specifically, we construct the matrix $\mathbf{B}$ by choosing the $K = \frac{G}{2}$ eigenvectors that describe the first-level detail coeffi-

cients. Therefore, $\mathbf{B}^\top\mathbf{B} = \mathbf{1}^K$ and its psuedoinverse is equal to its transpose $\mathbf{B}^+ = \mathbf{B}^\top$.

Given the basis $\mathbf{B}$ and the FIG prior in (4), it is easy to show that the prior distribution of the coefficients $\mathbf{C}$ can be expressed as

$$p(\mathbf{c}_f) = N\left(\mathbf{c}_f|\,\mathbf{0}, \sigma_f^2\mathbf{B}^\top\mathbf{D}_{\boldsymbol{\pi}_f}\mathbf{B}\right) \tag{7}$$

where σ_f^2 is further assumed to follow the *a priori* Inverse Gamma distribution

$$p(\sigma_f^2) = IG\left(\sigma_f^2\,|\,\alpha_f, \beta_f\right). \tag{8}$$

where α_f and β_f are respectively the shape and scale parameters, set up to have a non-informative prior [6] with $\alpha_n = 0.1$ and $\beta_n = 0.1$. On the other hand, the activities $\mathbf{x}_n$ and the noise variance σ_n^2 in (2) are modeled *a priori* by the conjugate Gaussian-Inverse-Gamma distribution as

$$p(\mathbf{x}_n, \sigma_n^2) = N\left(\mathbf{x}_n|\,0, \frac{\sigma_n^2}{\kappa_n}\mathbf{1}^F\right) IG\left(\sigma_n^2\,|\,\alpha_n, \beta_n\right) \tag{9}$$

where κ_n is a scale of the variance. To avoid scale unambiguity, this is set to be the inverse of the expected value of σ_n^2 with $\kappa_n = \frac{\beta_n}{\alpha_n+1}$. On the other hand, the shape and the scale parameters control the variance of the noise σ_n^2 and they are set to have a non-informative prior with $\alpha_n = 0.1$ and $\beta_n = 0.1$.

Given the gene expression data $\mathbf{Y}$ and the prior TF regulatory probabilities $\pi_{gf}\forall g, f$, the goal of uncovering TF mediated network is to infer the basis coefficients $\mathbf{C}$, from which the regulatory matrix $\mathbf{A}$ can be calculated, and the TF activities $\mathbf{X}$. We propose a Gibbs sampling solution in the next section.

3. GIBBS SAMPLING SOLUTION

A factor model considering the original loading matrix $\mathbf{A}$ with a prior Bernoulli Gaussian distribution demands the computation of all the G variables. In contrast, the BEFaM that we propose in (6) reduces the computational complexity to $K = \frac{G}{2}$ variables. Despite this reduction of the number of unknowns, the proposed model is still high-dimensional and the derivation of the posterior distributions is analytically intractable for a large number of genes. Therefore, we propose a Gibbs sampling solution. Note that σ_n^2 and σ_f^2 $\forall f, n$ are nuisance parameters to be estimated as well.

Gibbs sampling devises a Markov chain to produce random samples of the unknown from the intractable posterior distributions. The key of this method is to derive the conditional posterior distributions. Since all the priors are carefully chosen to be the conjugate priors, the close form of the marginal posterior distribution can be derived. Due to limited space, we omit the detailed derivation here. The Gibbs sampling draws samples from these marginal distributions iteratively and the tth iteration can be summarized as follows:

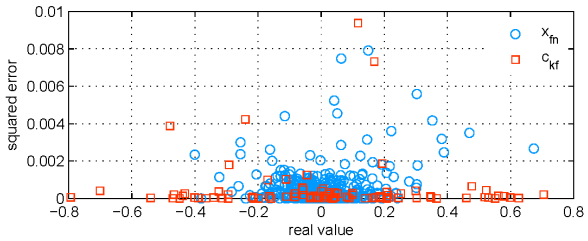


Fig. 2. Real values of the elements in $\mathbf{C}$ and $\mathbf{X}$ versus its squared error (SE) in the estimations.

- sampling $\hat{c}_{kf}^{(t+1)}$ from $p\left(c_{kf}|\,\mathbf{Y}, \left\{\hat{c}_{k\ell}^{(t)}\right\}_{\ell\neq f}^{F}, \mathbf{X}^{(t)}\right)$
- sampling $\hat{x}_{fn}^{(t+1)}$ from $p\left(x_{fn}|\,\mathbf{Y}, \left\{\hat{x}_{\ell n}^{(t)}\right\}_{\ell\neq f}^{F}, \mathbf{C}^{(t+1)}\right)$
- sampling $\hat{\sigma}_f^{2(t+1)}$ from $p\left(\sigma_f^2\middle|\,\mathbf{Y}, \hat{\mathbf{C}}^{(t+1)}, \hat{\mathbf{X}}^{(t+1)}\right)$
- sampling $\hat{\sigma}_n^{2(t+1)}$ from $p\left(\sigma_n^2\middle|\,\mathbf{Y}, \hat{\mathbf{C}}^{(t+1)}, \hat{\mathbf{X}}^{(t+1)}\right)$

Note that if the prior probabilities $\boldsymbol{\pi}_f$ in (7) are such that

$$\mathbf{b}_k^\top\mathbf{D}_{\boldsymbol{\pi}_f}\mathbf{b}_k = 0 \tag{10}$$

where $\mathbf{b}_k$ is the k-th eigenvector from the basis, then we have

$$p(c_{kf} = 0) = 1 \tag{11}$$

and no sampling is needed for c_{kf}.

To diagnose the convergence of Gibbs sampler, we adopt the scheme described in [6] by sampling parallel chains and discarding those corresponding to the burn-in period. Then, the rest of the samples of each chain are use to estimate the unknowns.

4. RESULTS

We have tested the proposed Gibbs sampling algorithm presented above using a simulated data, with $G = 50$ genes, $N = 50$ samples and $F = 8$ TFs. We generate a synthetic transcriptional network by the simulation of $\pi_{gf} \in [0.8, 1]$ for the 40% of its elements. The remaining 60% have a zero prior probability, a setting that mimics the real scenario as it is shown next. Subsequently, we generate the gene expression data set by simulations of the priors (7), (9) and (8) using the non-informative settings. As described above, the basis is built by choosing the first half eigenvectors from the Haar wavelet decomposition, with a dimension reduction of $K = \frac{G}{2} = 25$. The the Gibbs sampling considers $P = 10$ parallel chains, $T = 10000$ samples and a burning period of 5000 samples.

Figure 2 shows the results with the simulated data set as described above. This plot represent the normalized (non-zero) coefficients c_{kf} and the TF activities x_{fn} versus theirs

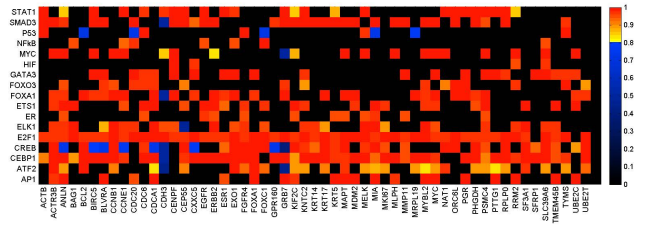


Fig. 3. Heatmap of the priors π_{gf} estimated by TransFact and MATCH, with proteins as rows and genes as columns.

squared errors (SE). The root mean squared error (RMSE) of the complete set of coefficients is $\text{RMSE}_{\mathbf{C}} = 4.75 \cdot 10^{-4}$, while for the TF activities is $\text{RMSE}_{\mathbf{X}} = 5.32 \cdot 10^{-4}$.

Besides the simulated data set, we have inferred the TF activities for a real data set with $N = 20$ breast cancer expression profiles whose subtypes are known: Basal and HER2. We consider a data set with $G = 55$ genes, the ones considered in the PAM50 test [7] that dissects the breast cancer in four main subtypes. Moreover, we consider $F = 17$ TFs by the appointment of experts and that are supposed to be related to breast cancer. Figure 3 represents the prior probabilities of the non-zero elements $\pi_{gf} \forall g, f$, with a 52% of sparsity level, predicted with the TransFac database and the Match tool [5] to perform TF binding site prediction. Figure 4 shows the hierarchical clustering resulting from the estimated TF activities. It is shown that the interfered protein profiles perfectly captures the biochemical differences between the two cancer subtypes. In the estimated transcriptional space, two proteins are highly correlated with the Basal and the HER2 breast cancer subtypes. The E2F1 and CEBP1, two proteins with a key role in the cell cycle and its apoptosis [8].

5. CONCLUSIONS

The BEFaM proposed in this paper constitutes a new way to model sparse networks to infer TF activities from gene expression data and gene-protein interaction prior knowledge. This new model introduce a FIG distribution and proposes a wavelet based expansion to reduce the complexity of the Gibbs sampling inference method. Our new method is validated by simulation and with real breast cancer data. Its performance shows satisfactory results with both simulated and real data, revealing its high potential in the disease molecular classification. As a future work, we propose to continue exploiting the the Bayesian formalism to combine data and to improve the solutions provided by the BEFaM.

6. REFERENCES

[1] Y. Huang, "Reverse engineering gene regulatory networks: A survey of statistical models," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 76–97, 2009.

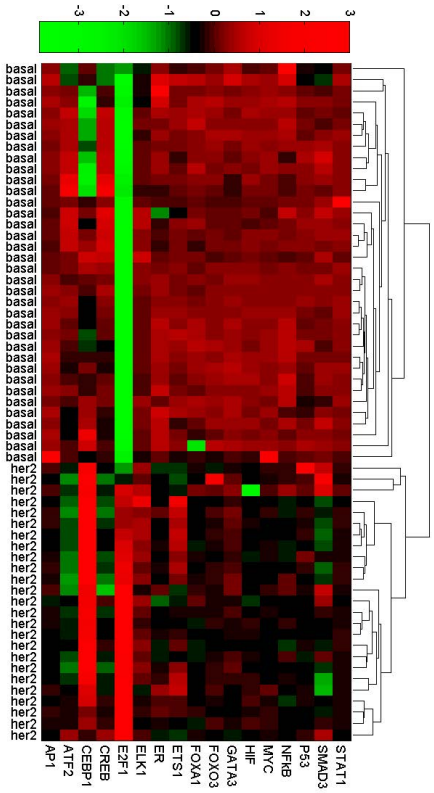


Fig. 4. Clustergram with the inferred protein activities for the breast cancer data.

[2] C. Sabatti, "Bayesian sparse hidden components analysis for TRN," *Bioinformatics*, vol. 22, 2006.

[3] J. Meng, "Uncovering Transcriptional Regulatory Networks by Sparse Bayesian Factor Model," *EURASIP Journal on Advances in Signal Processing*, 2010.

[4] C. Ye, "Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast," *PLoS Comp Bio*, vol. 5, pp. 1–12, 2009.

[5] V. Matys, "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Research*, vol. 34, pp. 108–110, 2006.

[6] A. Gelman, *Bayesian Data Analysis*, Chapman Hall.

[7] J. S. Parker, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinicical Oncology*, vol. 27, pp. 1160–1167, 2009.

[8] E. Mosca, "A multilevel data integration resource for breast cancer study," *BMC Systems Biology*, vol. 4, pp. 759–813, 2010.