A NEW OPTIMIZATION ALGORITHM FOR NETWORK COMPONENT ANALYSIS BASED ON CONVEX PROGRAMMING

Chunqi Chang[1], *Yeung Sam Hung*[1]

Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong

Zhi Ding

Electrical and Computer Engineering
University of California
Davis, CA95616, USA

ABSTRACT

Network component analysis (NCA) has been established as a promising tool for reconstructing gene regulatory networks from microarray data. NCA is a method that can resolve the problem of blind source separation when the mixing matrix instead has a known sparse structure despite the correlation among the source signals. The original NCA algorithm relies on alternating least squares (ALS) and suffers from local convergence as well as slow convergence. In this paper, we develop new and more robust NCA algorithms by incorporating additional signal constraints. In particular, we introduce the biologically sound constraints that all nonzero entries in the connectivity network are positive. Our new approach formulates a convex optimization problem which can be solved efficiently and effectively by fast convex programming algorithms. We verify the effectiveness and robustness of our new approach using simulations and gene regulatory network reconstruction from experimental yeast cell cycle microarray data.

Index Terms— Gene regulatory networks, microarray, network component analysis, convex programming

1. INTRODUCTION

Recent advances in genome sequencing and high-throughput microarray technologies make it possible to quantitatively investigate the underlying mechanisms that define gene interaction. More specifically, we can now consider the underlying biological model as an information processing system. With this signal processing framework, the challenge is to model gene interaction from experimental microarray data as a gene regulatory network. A key focus is on developing tractable system identification techniques capable of reconstructing gene regulatory networks of a large size from small microarray datasets.

Earlier approaches to modeling gene regulatory network include dynamic Bayesian networks (DBN) [1], probabilistic Boolean network (PBN) [2], and differential/difference equations [3]. It has also been shown that the network can be approximated by a linear instantaneous signal system [4]. Among various algorithms assuming linear instantaneous signal models, such as principal component analysis (PCA) [5] and independent component analysis (ICA)[6], the network component analysis (NCA)[4] is a very effective approach since it incorporates useful and biologically sound assumptions.

More specifically, a true gene regulatory network always has a sparse structure, and sometimes the non-zero entries of its (sparse) connectivity matrix can be obtained from ChIP-chip experiments [7]. Even when experimental data on the structure of the network are unavailable, the structural information sometimes can be extracted partially from the literature or predicted by bioinformatics methods [8]. The NCA approach makes use of this (sparse) structure information. If some mild conditions, called NCA criteria, can be satisfied, the NCA algorithm can fit the gene expression data (measured by microarray) to the linear network model under the constraints represented by the NCA criteria. The NCA algorithm can then fully reconstruct the network (including the connectivity matrix and the regulatory signals) if the measurement is noise-free.

The original NCA algorithm applies alternating least squares (ALS) to solve a not very well posed optimization problem. Though very effective, it suffers from three computational drawbacks as pointed out in [9]:

(1) it may be unstable due to ill-conditioned matrices;
(2) it may converge to local minima;
(3) it is inefficient and time consuming for relatively large networks.

Tikhov regularization has been proposed to overcome the convergence problem [9]. Most recently, authors of this manuscript proposed a new algorithm, FastNCA, that overcome the 3 drawbacks [10]. FastNCA provides a closed-form solution to NCA through matrix factorizations.

One common feature of the existing NCA approaches is that it exploit only the knowledge on the zero location of the (sparse) connectivity matrix. Since microarray data are very

[1] Supported by The University of Hong Kong CRCG grants.

noisy, we seek to incorporate additional constraints to reduce noise sensitivity. From biological information, we find that, without loss of generality, all nonzero entries in the connectivity matrix can be positive. We hence develop a new NCA algorithm to take advantage of such constraints against noise. We further formulate an optimization problem that can be solved by efficient convex programming algorithms.

Our paper is organized as follows. In Section 2 we introduce the model of gene regulatory network and the NCA formulation. We formulate, in Section 3, a convex optimization problem to tackle the NCA problem. In Section 4 we impose new solution constraints onto the NCA problem, and incorporate efficiently in the convex programming framework. We provide simulation and experimental results in Section 5, followed by discussions and conclusions in Section 6.

2. GENE REGULATION MODEL AND NETWORK COMPONENT ANALYSIS

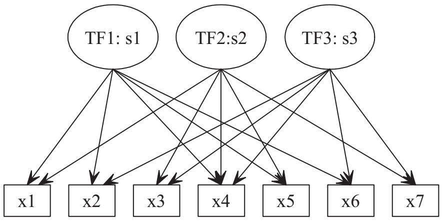


Fig. 1. Transcriptional regulatory model.

We consider gene expression with a transcriptional regulatory model, as illustrated by Figure 1. Gene expressions are regulated by transcription factors. The upper layer in the figure represents the expression level of activated transcription factors (TF) or transcription factor activities (TFA), and the lower layer represents the microarray gene expression data. Consider a network of N genes regulated by M TFs with observed measurements over K time points. This network can be modeled as an instantaneous linear mixing system

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{\Gamma}, \tag{1}$$

where the mixtures $\mathbf{X}$ (with dimension $N \times K$) are gene expression levels which can be measured by microarray, the mixing matrix $\mathbf{A}$ ($N \times M$) is called connectivity matrix of the network, the sources $\mathbf{S}$ ($M \times K$) are the unknown TFAs, and $\mathbf{\Gamma}$ is the measurement noise.

We only have access to the gene expression data X, i.e., the microarray data. The aim is to estimate the connectivity matrix $\mathbf{A}$ and the TFAs $\mathbf{S}$ from the microarray data $\mathbf{X}$. Our strategy is to estimate the connectivity matrix first. With the estimation of $\mathbf{A}$, denoted by $\hat{\mathbf{A}}$, from which the TFAs can be estimated by algorithms such as minimum mean square error (MMSE) or zeroforcing inverse from

$$\hat{\mathbf{S}} = \hat{\mathbf{A}}^{+}\mathbf{X}, \tag{2}$$

where $\hat{\mathbf{A}}^{+}$ is the pseudo inverse of $\hat{\mathbf{A}}$.

The method developed in [4] to solve such inverse problem is called network component analysis (NCA), which is based on the following three assumptions referred to as *NCA criteria*: (i) connectivity matrix $\mathbf{A}$ must be full-column rank; (ii) when an element in the regulatory domain is removed along with all the output elements connected to it, the connectivity matrix of the resulting network is still of full-column rank; and (iii) $\mathbf{S}$ must have full row rank.

3. A CONVEX PROGRAMMING NCA FRAMEWORK

Here we develop a new approach to network component analysis, which is based on the same three NCA criteria but, like our FastNCA approach, does not suffer the drawbacks of the original NCA algorithm using ALS and its improved version with regularization. In addition, this approach, as will be shown later, can easily incorporate other kinds of prior information to achieve improved performance.

First we can find the range of $\mathbf{X}$: $\bar{\mathbf{X}} = \text{range}\{\mathbf{X}\}$, and its null space $\mathbf{C} = \bar{\mathbf{X}}^{\perp}$ such that $\mathbf{C}^T\bar{\mathbf{X}} = 0$. When the measurement is noiseless, i.e., $\mathbf{X} = \mathbf{A}\mathbf{S}$, then $\bar{\mathbf{X}}$ and $\mathbf{C}$ are also the range and null space of $\mathbf{A}$. When there is noise, we can use a robust estimation of $\mathbf{C}$ through singular value decomposition (SVD) [11]:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{3}$$

in which the diagonal matrix Σ consists of singular values sorted in descending order. From which we can estimate $\mathbf{C}$ as the last $N - M$ columns of $\mathbf{U}$.

Because $\mathbf{A}$ is orthogonal to $\mathbf{C}$, this condition can be used to estimate $\mathbf{A}$ by minimizing the Frobenius norm $\|\mathbf{C}^T\mathbf{A}\|_F$ subject to the constraints imposed by the prior information on network connectivity. In other words, we propose to estimate $\hat{\mathbf{A}}$ as the solution of the following optimization problem.

$$\min_{\hat{\mathbf{A}}} ||\mathbf{C}^T\hat{\mathbf{A}}||_F \quad \text{subject to} \quad \hat{\mathbf{A}}(I) = 0, \tag{4}$$

where I contains the indices where the entries of $\mathbf{A}$ are zeros.

This optimization problem is globally convergent. First, the cost function is a convex function of the unknown matrix $\hat{\mathbf{A}}$. Second, the constraints $\hat{\mathbf{A}}(I) = 0$ are linear (hence convex). Thus, the optimization problem 4 is a convex programming problem and contains no local minima. It can also be solved with very fast interior-point algorithms [12].

4. POSITIVITY CONSTRAINTS

Though the three NCA criteria are enough to estimate the connectivity matrix when there is no noise in the model, additional constraints, if can be used in the algorithm, will certainly yield a more robust estimate in practice when noise is inevitable. The original approach in [9] is not easily amenable to the addition of additional constraints and prior knowledges. Our convex optimization approach, on the other hand, is very simple for integration with new (convex) constraints and can often converge faster. This property makes it perfectly suited to integrate other system knowledge to combat measurement noises and modeling errors.

In this paper, we assume the prior knowledge that all nonzero entries of the connectivity matrix are of the same sign (either all positive or all negative). As long as any specific transcription factor has the same effect to all genes, either all positive or all negative, this assumption has sound biological support [13]. This is because if a transcription factor regulates the genes negatively, then we can simply multiply its transcription factor activity (TFA) by -1 to meet the requirement that the new TFA will have positive effect on the genes.

Note that the positivity constraints are linear inequalities and are convex, they are easily incorporated into a modified convex programming for NCA (named PosNCA) as

$$\min_{\hat{\mathbf{A}}} ||\mathbf{C}^T \hat{\mathbf{A}}||_F \quad \text{subject to} \quad \hat{\mathbf{A}}(I) = 0,\ \hat{\mathbf{A}}(J) \geq c, \qquad (5)$$

where J contains the indices where entries of $\mathbf{A}$ are nonzero, and c is a constant small positive value. The convex problem 5 can be solved by effective algorithms [12].

5. RESULTS

5.1. Simulation results

Without positivity constraints, simulations (not shown here) demonstrated that the convex programming of (4) matches well with the results of FastNCA.

The new PosNCA with positivity constraints is tested on the chemical simulation data identical to [4]. This is a hemoglobin spectroscopy data set with a 7×3 mixing matrix and a 7×321 measurement.

Without noise, both the source spectra and the mixing matrix can be perfectly estimated using convex programming with positivity constraints. Here we study the proposed algorithm under noisy case. The FastNCA method [10] is also used as a comparison. To this end we add noise to the spectral measurement to get a signal to noise ratio (SNR) of 9dB. The true (original) and the estimated source spectra are shown in in Figure 2, while the (normalized) true and estimated mixing matrices are compared in Table 1.

The estimated entries of $\mathbf{A}$ by FastNCA can be both positive and negative. Such estimate is unfavorable since all true entries in $\mathbf{A}$ are non-negative. On the other hand, the estimated connectivity matrix by the PosNCA algorithm contains much closer results as expected. The root mean square error of the $\mathbf{A}$ estimate from PosNCA is 0.3114, much smaller than that from FastNCA (0.7328). The estimated source spectra shown in Figure 2 also demonstrate that PosNCA is much more robust.

Table 1. True connectivity matrix and its estimation by FastNCA and the proposed PosNCA algorithm

True			FastNCA			PosNCA		
0.417	1.18	0	-1.78	1.03	0	0.985	1.03	0
2.08	0	1.25	-0.20	0	2.06	1.51	0	1.24
0	1.18	0.25	0	1.24	-0.47	0	1.24	0.670
0.417	1.05	0.25	-0.946	0.966	0.096	1.03	0.969	0.670
1.25	0.921	0	-0.279	1.09	0	0.741	1.09	0
0.833	0	2	1.79	0	1.78	0.741	0	1.75
0	0.658	1.25	0	0.675	0.589	0	0.677	0.670

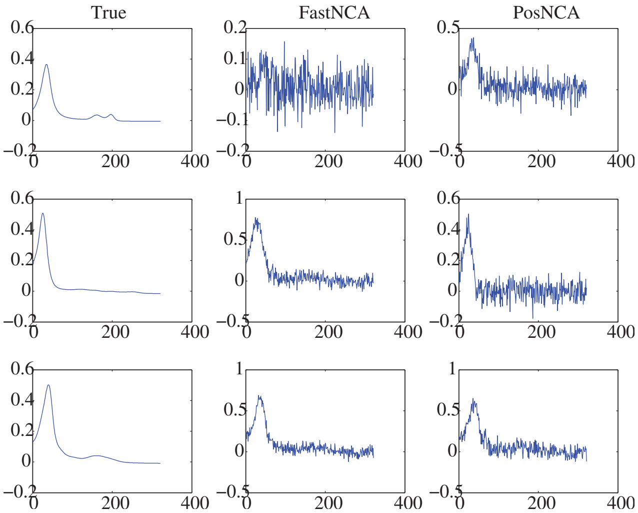
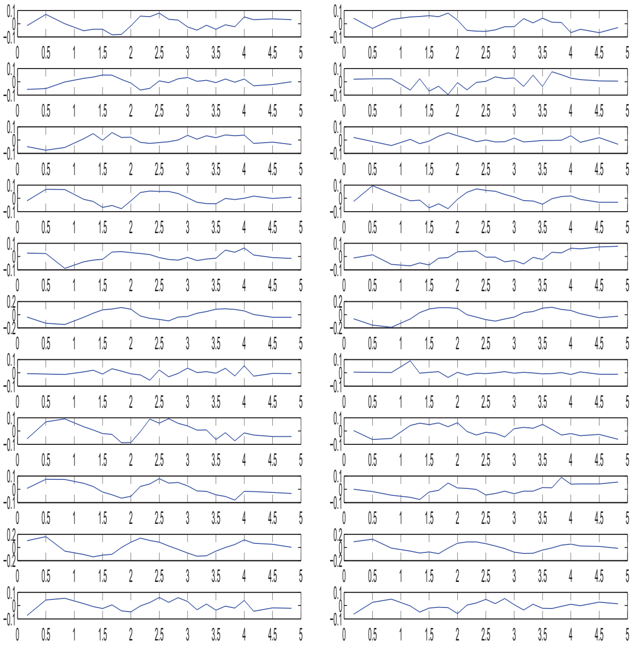


Fig. 2. True sources and their estimates using FastNCA and PosNCA.

5.2. Experimental Results

The time-course Yeast cell cycle microarray data from [14] are analyzed by our PosNCA with positivity constraints. The network topology data are from [7]. After trimming, the network has 441 genes and 33 transcription factors in which 11 TFs are cell cycle related. The estimated TFAs for these 11 cell cycle related transcription factors are shown in Figure 3 for the new PosNCA algorithm and the previous FastNCA algorithm, respectively. Only results from the experiment using the synchronization method by arrest of cdc15 temperature-sensitive mutant are presented here. It can be seen that the two methods produce similar results in general. However, it is clear that the estimated TFAs by the new PosNCA exhibit more cyclic behaviors. Cyclic behavior is expected in this case since all these transcription factors are cell cycle related. Therefore, it leads to the reasonable speculation that the PosNCA results are more reliable than those from other NCA methods without incorporating these biologically sound constraints. Still, we stress that such conclusion is to be confirmed by further biological studies.



(a) PosNCA (b) FastNCA

Fig. 3. Estimated yeast cell cycle related TFAs by (a) the proposed PosNCA method; and (b) FastNCA

6. DISCUSSION AND CONCLUSION

According to the biological literatures, the microarray data are in fact considered very noisy. Thus, the development of PosNCA with positivity constraints is expected to provide practical advantage against measurement noise by incorporating biologically sound constraints. The constraints are seamlessly integrated into our convex optimization algorithm. As shown by the yeast cell cycle data analysis, the new PosNCA algorithm with positivity constraints presents better results than FastNCA without such constraints.

7. REFERENCES

[1] N. Friedman, M. Linial, Nachman I., and Peer D., "Using bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[2] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[3] E. R. Dougherty, I Shmulevich, J Chen, and Z. J. Wang, *Genomic Signal Processing and Statistics*, Hindawi Publishing Corporation, 2005.

[4] J C Liao, R Boscolo, Y L Yang, L M Tran, C Sabatti, and V P Roychowdhury, "Network component analysis: Reconstruction of regulatory signals in biological systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 26, pp. 15522–15527, 2003.

[5] O Alter, P O Brown, and D Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 18, pp. 10101–10106, 2000.

[6] S I Lee and S Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology*, vol. 4, no. 11, pp. R76, 2003.

[7] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young, "Transcriptional regulatory networks in saccharomyces cerevisiae," *Science*, vol. 298, no. 5594, pp. 799–804, 2002.

[8] C. Wang, J. Xuan, L. Chen, P. Zhao, Y. Wang, R. Clarke, and E. Hoffman, "Motif-directed network component analysis for regulatory network inference," *BMC Bioinformatics*, vol. 9, no. Suppl 1, pp. S21, 2008.

[9] L M Tran, M P Brynildsen, K C Kao, J K Suen, and J C Liao, "gnca: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation," *Metabolic Engineering*, vol. 7, no. 2, pp. 128–141, 2005.

[10] C. Q. Chang, Z. Ding, Y. S. Hung, and P. C. W. Fung, "Fast network component analysis (fastnca) for gene regulatory network reconstruction from microarray data," *Bioinformatics*, vol. 24, pp. 1349–1358, 2008.

[11] G. H. Golub and C. F. van Loan, *Matrix Computation*, The Johns Hopkins University Press, 3rd edition, 1996.

[12] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, UK, 2004.

[13] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC, 2007.

[14] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, Botsein D., and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.