GENE DELETION DATA BASED GENOMIC REGULATORY NETWORK INFERENCE

Liming Wang and Xiaodong Wang

Department of Electrical Engineering
Columbia University, New York, NY 10027
{liming, wangx}@ee.columbia.edu

ABSTRACT

The gene deletion data is a type of gene expression data, which is obtained by deleting each gene consecutively from the network and measuring the fitness of the remaining network under various environmental conditions. Compared to the microarray data, the deletion data is much easier and economical to obtain. The gene tag technology has enabled the deletion data to be largely available for various regulatory networks. However, very few inference algorithms are proposed for the deletion data in spite of its advantages. In this paper, we propose an inference algorithm based on gene deletion data. The proposed inference algorithm capture the dynamical and non-linear natures of the regulatory networks. We conduct experiment on the *GAL* network to demonstrate the performance of the proposed algorithm. The proposed algorithm has been shown to serve as a good alternatives for exploring various regulatory networks other than using microarray data.

Index Terms— Gene deletion, unscented Kalman filter, microarray

1. INTRODUCTION

The inference of gene regulatory networks have become an attracting research area in recent years thanks to the availability of DNA microarray technology [1]. The vast amount of data obtained by the microarray technology enables the possibility of accurate estimation of gene regulatory network structures, which has been proven to be an important basis for medical diagnosis and treatment. It is well-known that gene expressions are inherently stochastic. Therefore the expressions for genes can be viewed as stochastic time-series data. The goal of the inference algorithm is to discover the connectivity structure based on these time-series data.

Inference algorithms vary for different modelings of the network. There are a number of modelings proposed for the regulatory networks [1, 2, 3]. One category of these models quantizes the expressions to binary numbers and views the structures of network as Boolean constraints. Another category considers the network in continuous time with introducing the differential equations. Beside of these approaches, the models from control and stochastic differential equation point of view are also popular, which include the state-space and stochastic model.

The cost of microarray technology has reduced significantly in the past few years with the advancement of technology. However, current experimental cost still hinders the possibility of acquiring sufficient data especially for large-scale genome networks. Meanwhile, the idea of deleting genes sequentially from the network and its experimental technique have attracted many attentions recently [4]. The gene deletion data usually measures certain factor such as growth rate under various experimental conditions after sequentially deleting one gene from the network. Compared to traditional microarray technology, the deletion data is largely available for various networks such as the yeast *Saccharomyces cerevisiae*.

Very few systematical models and inference algorithms are proposed for the regulatory networks via the deletion data, to our best of knowledge. Some results are obtained by visual inspections or relatively naive strategies. As we mentioned before, because of the increasing popularity of deletion data, an inference algorithm utilizing the deletion data is needed. The inference result can be used as an alternative way for understanding regulatory networks. It can also be employed to find a good initialization for further inference using the valuable and limited microarray data, whose performance can be significantly improved by starting from a good initialization.

In this paper, we propose a dynamic model and inference algorithm for gene regulatory networks based on the gene deletion data. We resort to the unscented Kalman filter (UKF) approach to estimate all the parameters. We also provide experimental results for the proposed algorithm on the *Saccharomyces cerevisiae* data. We have compared our inferred results to some known facts for justification of the accuracy.

2. SYSTEM MODEL

We first provide a model for microarray data, on which we will derive the system model under gene deletion data. Consider a gene regulatory network with total N genes. Let $g_i(k)$, $i = 1, \ldots, N$, $k = 1, 2, \ldots, M$ denote the gene expression level for the i-th gene at time k. For observation or measure-

ment data $x_i(k)$ for $g_i(k)$ at time k, we model it as

$$x_i(k) = g_i(k) + v_i(k), \tag{1}$$

where $v_i(k)$ is the observation noise at time k for i-th gene. We denote all the expression levels of the network as a vector $\mathbf{g}(k) = [g_1(k), \ldots, g_N(k)]^T$, the observation vector as $\mathbf{x}(k) = [x_1(k), \ldots, x_N(k)]^T$ and noise vector as $\mathbf{v}(\mathbf{k}) = [v_1(k), \ldots, v_N(k)]^T$. We assume all the vectors $\mathbf{v}(k)$ for $k = 1, \ldots, M$ are indepedent and jointly Gaussian with zero mean and variance matrix $\mathbf{R}(k)$.

We follow the discrete-time gene regulation model proposed in [5], where the regulatory functions among all the genes are

$$g_i(k+1) = \sum_{j=1}^{N} a_{ij} g_j(k) + \sum_{j=1}^{N} b_{ij} f_j(g_j(k), \mu_j) + I_i + w_i(k), \tag{2}$$

for $i = 1, \ldots, N$, where a_{ij} denotes the linear regulation coefficient from gene j to gene i; b_{ij} denotes the non-linear regulation coefficient from gene j to gene i; f_j is the non-linear function for gene j which is given by

$$f_j(g_j, \mu_j) = \frac{1}{1 + e^{-\mu_j g_j}}, \tag{3}$$

μ_j is the parameter to be inferred; I_i denotes the system expression bias for i-th gene which will be inferred later. The noise vectors $\mathbf{w}(k) = [w_1(k), \ldots, w_N(k)]^T$ for $k = 1, 2, \ldots, M$ are assume to be jointly Gaussian with zero mean and variance $\mathbf{Q}(k)$. We also assume that they are independent to all $\mathbf{V}(k)$. We denote $\mathbf{A} = [a_{11}, a_{12}, \ldots, a_{NN}]^T$; $\mathbf{B} = [b_{11}, b_{12}, \ldots, b_{NN}]^T$; $\mathbf{I} = [I_1, \ldots, I_N]^T$ and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]^T$.

The goal for inference is to estimate all the unknown parameters in the model. The equations (1) and (2) determine the dynamic of the system for acquisition of the microarray data. However, necessary changes are required to model the acquisition of deletion data. After we fix the labeling of indices of the genes, we assume that we delete the gene sequentially from 1 to N. At time step k, we delete gene with index (k mod N) if $k \nmid N$, or N otherwise.

Without loss of generality, we assume that when the system is at time k, j_k-th gene has been deleted. Note that the index j_k is determined as we mentioned in previous paragraph. For the state-space model, all the gene expressions evolve without participation of gene j_k. Therefore we have all the regulatory coefficients $a_{\cdot j_k}(k) = 0$ and $b_{\cdot j_k}(k) = 0$. In other words, we view the system as a time-variant system. The state of gene j_k should remain unchanged since it has been deleted from the network. The states and system coefficients equations can be summarized as

$$\begin{aligned}
I_i(k+1) &= I_i(k) \quad \forall i, \\
\mu_i(k+1) &= \mu_i(k) \quad \forall i, \\
a_{i,j_k}(k) &= 0 \quad \forall i, \\
b_{i,j_k}(k) &= 0 \quad \forall i, \\
a_{i,j}(k+1) &= a_{i,j}(k), \quad \text{if } j \neq j_k, \\
b_{i,j}(k+1) &= b_{i,j}(k), \quad \text{if } j \neq j_k, \\
g_i(k+1) &= \sum_{j=1}^{N} a_{ij}(k) g_j(k) + \sum_{j=1}^{N} b_{ij}(k) f_j(g_j(k), \mu_j) \\
&\quad + I_i + w_i(k), \quad \text{if } i \neq j_k, \\
g_{j_k}(k+1) &= g_{j_k}(k).
\end{aligned} \tag{4}$$

The last two equations determine the current system coefficients, which will be augmented into the system states. Unlike the case in microarray measurements, the measurements in deletion datasets are obtained by measuring a factor which is a function of all the remaining genes. In this paper, we assume the measurement is a real number, which represents the fitness of the remaining network. The model can be easily adapted into higher dimensional measurement case. Therefore the observation $x(k)$ is

$$x(k) = f(g_1(k), \ldots, g_N(k)) + V(k), \tag{5}$$

where $f : \mathbb{R}^N \to \mathbb{R}$ is the experimental function which usually is not known *a priori*. We denote $R(k)$ as the variance of $V(k)$ as before. In order to estimate f, we will use various basis to approximate it and augment the states equations for estimation of the coefficients associated to those basis. The radial basis approach has been shown to be more robust and adaptive than Taylor's expansion.

More specifically, we approximate f as

$$f(\mathbf{y}) \approx \sum_{j=1}^{p} \lambda_j \Phi(\|\mathbf{y} - \mathbf{y}_j^p\|) + \boldsymbol{\lambda}_0^T \mathbf{y}, \tag{6}$$

where λ_i for $i = 1, \ldots, p$ and $\boldsymbol{\lambda}_0$ are the centers of the basis; p is the total number of basis which is a fixed constant; $\Phi(x) := \sqrt{c + x^2}$ is the Hardy multi-quadratic function, where $c > 0$ is a constant. Let $\boldsymbol{\lambda} = [\boldsymbol{\lambda}_0^T, \lambda_1, \ldots, \lambda_p]^T$. All the coefficients $\lambda_i, i = 1, \ldots, p$ and $\boldsymbol{\lambda}_0$ are parameters to be inferred which will be combined into the state variables. Therefore, the new augmented state variable is

$$\mathbf{y}(k) = [\mathbf{g}^T(k), \mathbf{A}^T, \mathbf{B}^T, \mathbf{I}^T, \boldsymbol{\mu}^T, \boldsymbol{\lambda}^T]^T \tag{7}$$

As a summary, the dynamical model we propose for regulatory network under the gene deletion data is

$$\begin{aligned}
\mathbf{y}(k+1) &= F_k(\mathbf{y}(k)) + \mathbf{W}(k) \\
x(k) &= f(I_k \mathbf{y}(k)) + V(k),
\end{aligned} \tag{8}$$

where $F_k(\cdot)$ is the system function described in (4); $\mathbf{W}(k) = [w_1, \ldots, w_N, 0, \ldots, 0]$ is the augmented noise vector; I_k is the selection matrix, i.e.,
$I_k\mathbf{y} = [g_1, \ldots, g_{j_\star - 1}, 0, g_{j_\star + 1}, \ldots, g_N]^T$, where the index $\star$ is determined by the time index k; $V(k)$ is the Gaussian noise as assumed before.

3. THE UNSCENTED KALMAN FILTER APPROACH FOR INFERENCE

In previous section, we propose the system model for the regulatory network based on the state-space model. In order to infer all the unknown parameters, we utilize the UKF approach [6]. UKF enjoys many advantages to the classical extended Kalman filter (EKF) approach. The UKF is based on the idea choosing sigma points from the unscented transform. For a random vector $\mathbf{x}$, let us consider it passing through a non-linear transform $\mathbf{y} = h(\mathbf{x})$. In order to calculate the mean and variance of $\mathbf{y}$, we choose the sigma points $\mathcal{S}_i$, $i = 0, \ldots, 2R$ and their weights $\mathcal{W}_i$ as follow,

$$\begin{aligned}
\mathcal{S}_0 &= E(\mathbf{x}), \\
\mathcal{S}_i &= E(\mathbf{x}) + (\sqrt{(L+\lambda)\mathrm{Var}(\mathbf{x})})_i \quad i = 1, \ldots, R, \\
\mathcal{S}_i &= E(\mathbf{x}) - (\sqrt{(L+\lambda)\mathrm{Var}(\mathbf{x})})_{i-R} \quad i = R+1, \ldots, 2R, \\
\mathcal{W}_0^{(m)} &= \frac{\lambda}{L+\lambda}, \\
\mathcal{W}_0^{(p)} &= \frac{\lambda}{(L+\lambda)} + (1 - \alpha^2 + \beta), \\
\mathcal{W}_i^{(m)} &= \mathcal{W}_i^{(p)} = \frac{1}{2(L+\lambda)} \quad i = 1, \ldots, 2R,
\end{aligned} \tag{9}$$

where $\mathrm{Var}(\mathbf{x})$ is the variance matrix of random variable $\mathbf{x}$; $(\cdot)_i$ denotes the i-th column of the input matrix; $\lambda = \alpha^2(R + \kappa) - R$ is the scaling parameter. β is a parameter to incorporate prior knowledge of $\mathbf{x}$. Since we have the Gaussian noise assumption, we choose $\kappa = 0$, $\beta = 2$ and $\alpha = 10^{-3}$. We refer readers to [6] for specific choices of all these parameters.

In order to infer the model (8), we simply concatenate the state variable $\mathbf{y}$ with the noise vectors $\mathbf{W}$ and V to form a new augmented vector

$$\mathbf{y}^a(k) = [\mathbf{y}^T(k), \mathbf{W}^T(k), V(k)]^T. \tag{10}$$

We view the $F_k(\cdot)$ and $f(\cdot)$ as the non-linear transforms and calculate the sigma points to approximate the mean and variance which will be used in the sequential updates. As we see here, another advantage to the EKF approach is that we do not require calculate the Jacobian or Hessian of the transform, which makes the algorithm and mathematical derivations less involved.

Now we provide the UKF based inference algorithm for model (8) based on gene deletion data.

- Initialize the state variable with

$$\begin{aligned}
\hat{\mathbf{y}}(0) &= \mathbf{0}, \\
\mathbf{P}(0) &= I, \\
\mathbf{y}^a(0) &= [\hat{\mathbf{y}}^T(0), \mathbf{0}, 0]^T, \\
\mathrm{Var}(\hat{\mathbf{y}}^a(0)) &= \begin{pmatrix} \mathbf{P}(0) & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{Q}(0) & 0 \\ \mathbf{0} & \mathbf{0} & R(0) \end{pmatrix},
\end{aligned} \tag{11}$$

  where I is the identity matrix.

- For time step $k = 1, 2, \ldots$, calculate the sigma points $\{\mathcal{S}^y_{k-1}, \mathcal{S}^W_{k-1}, \mathcal{S}^v_{k-1}\}$ for $\hat{\mathbf{y}}(k-1)$, $\mathbf{W}(k-1)$ and $V(k-1)$ respectively by (9), where $\{\mathcal{S}^y_{k-1}, \mathcal{S}^W_{k-1}, \mathcal{S}^v_{k-1}\}$ denote the sigma matrices, each of which is obtained by combining all the corresponding sigma point vectors as columns together.

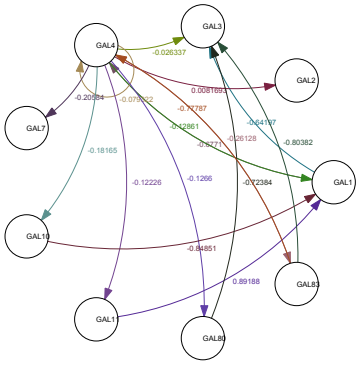
- Time update equations:

$$\begin{aligned}
&\mathcal{S}^y_i(k|k-1) = F_k((\mathcal{S}^y_{k-1})_i) + (\mathcal{S}^W_{k-1})_i \quad i = 0, \ldots, 2R, \\
&E^y_S(k) = \sum_{i=0}^{2R} \mathcal{W}_i^{(m)} \mathcal{S}^y_i(k|k-1), \\
&\mathrm{Var}^y_S(k) = \\
&\sum_{i=0}^{2R} \mathcal{W}_i^{(c)} (\mathcal{S}^y_i(k|k-1) - E^y_S(k))(\mathcal{S}^y_i(k|k-1) - E^y_S(k))^T, \\
&\mathcal{S}^x_i(k|k-1) = \\
&f(I_k(\mathcal{S}^y_i(k|k-1))_i) + (\mathcal{S}^V_{k-1})_i \quad i = 0, \ldots, 2R, \\
&E^x_S(k) = \sum_{i=0}^{2R} \mathcal{W}_i^{(m)} \mathcal{S}^x_i(k|k-1).
\end{aligned} \tag{12}$$

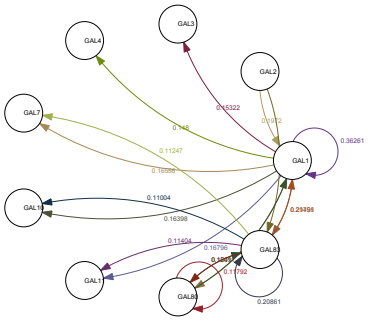
  In above equations, $(\cdot)_i$ denotes the i-th column of the input matrix as before.

- Measurement update equations:

$$\begin{aligned}
&\mathrm{Var}^x_S(k) = \\
&\sum_{i=0}^{2R} \mathcal{W}_i^{(c)} (\mathcal{S}^x_i(k|k-1) - E^x_S(k))(\mathcal{S}^x_i(k|k-1) - E^x_S(k))^T, \\
&\mathrm{Var}^{yx}_S = \\
&\sum_{i=0}^{2R} \mathcal{W}_i^{(c)} (\mathcal{S}^y_i(k|k-1) - E^y_S(k))(\mathcal{S}^x_i(k|k-1) - E^x_S(k))^T, \\
&\mathcal{K} = \mathrm{Var}^{yx}_S(k) \mathrm{Var}^x_S(k)^{-1}, \\
&\hat{\mathbf{y}}(k) = E^y_S(k) + \mathcal{K}(x(k) - E^x_S(k)), \\
&\mathbf{P}(k) = \mathrm{Var}^y_S(k) - \mathcal{K}\mathrm{Var}^x_S(k)\mathcal{K}^T, \\
&\mathrm{Var}(\hat{\mathbf{y}}^a(k)) = \begin{pmatrix} \mathbf{P}(k) & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{Q}(k) & 0 \\ \mathbf{0} & \mathbf{0} & R(k) \end{pmatrix}.
\end{aligned} \tag{13}$$



(a) Structure of the inferred network based on linear connections.



(b) Structure of the inferred network based on non-linear connections.

Fig. 1. Inference result for *GAL* regulatory network.

4. EXPERIMENTS

We apply the proposed inference algorithm to the *GAL* regulatory network. The *GAL* families are collection of genes which control the utilization of galactose in yeast *Saccharomyces cerevisiae*. We use the gene deletion data from [7]. The one dimensional measurements are measured under different environmental conditions such as different concentrations of galactose, Alkali, Sodium chloride, Sorbitol. etc.

In Fig. 1, we show the inferred network structures for linear and non-linear coefficients respectively. In order to evaluate the performance of the inference algorithm, we compare our results to various known facts about the *GAL* network [8].

First it can be seen that from Fig. 1, *GAL 1, 3, 4* and 80 globally have the most connectivities and largest connection coefficients, which are in accordance with known facts that they are the regulation genes in the network with other genes regarded as the structural genes. Besides, we see that *GAL 80* has a negative connection to *GAL 3* and *4*; *GAL 4* has negative connection to *GAL 1* and *7*, all of which coincide with the facts that *GAL 80* has negative regulations on *GAL 3, 4* and *GAL 4* prevents transcriptions of *GAL 1, 7*.

Moreover, the connections between *GAL 1, 2* and *GAL 3* reflect the fact that *GAL 2* and *GAL 1* regulate *GAL 3* by protein utilization pathway. We see that there is no direct connection from *GAL 11* to *GAL 80* which also coincides with the fact that *GAL 11* does not have direct interaction to *GAL 80*.

We find the inference results miss or contradict to some known facts about the network. However, most of these inconsistencies are among the structural genes. The inferred regulations among regulatory genes are fairly accurate. Because the inherent under-determined nature of this inference problem may compromise the performance of inference. We probably should not expect that the quality of inference can match the inference via the microarray data. On the other hand, the ampleness of data may remise that problem to some extent. Nevertheless, we see that the proposed inference algorithm provides a fairly satisfying result for the *GAL* network.

5. REFERENCES

[1] H. de Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *J. of Comp. Bio.*, vol. 9, no. 1, pp. 67–103, 2002.

[2] I. Shmulevich, E.R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," *Proc. of the IEEE*, vol. 90, no. 11, pp. 1778–1792, 2002.

[3] L. Wang and D. Schonfeld, "Game theoretic model for control of gene regulatory networks," in *Proc. of IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE, 2010, pp. 542–545.

[4] J. Oh et al, "A universal TagModule collection for parallel genetic analysis of microorganisms," *Nucleic Acids Research*, vol. 38, no. 14, pp. e146, 2010.

[5] L. Chen and K. Aihara, "Chaos and asymptotical stability in discrete-time neural networks," *Physica D: Nonlinear Phenomena*, vol. 104, no. 3-4, pp. 286–325, 1997.

[6] S.J. Julier and J.K. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," in *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*. Spie Bellingham, WA, 1997, vol. 3, pp. 26–37.

[7] G. Giaever et al, "Functional profiling of the Saccharomyces cerevisiae genome," *Nature*, vol. 418, no. 6896, pp. 387–391, 2002.

[8] D. Lohr, P. Venkov, and J. Zlatanova, "Transcriptional regulation in the yeast GAL gene family: a complex genetic network," *The FASEB Journal*, vol. 9, no. 9, pp. 777–787, 1995.