INFERRING GENE REGULATORY NETWORKS WITH NONLINEAR MODELS VIA EXPLOITING SPARSITY[‡]

Amina Noor, Erchin Serpedin[*]

Texas A&M University
Department of ECE
College Station, Texas 77843-3128

Mohamed Nounou, Hazem Nounou

Texas A&M University at Qatar
Doha, Qatar

ABSTRACT

This paper considers the problem of inferring gene regulatory networks using time series data. A nonlinear model is assumed for the gene expression profiles, whereas the microarray data follows a linear Gaussian model. A particle filter based approach is proposed to estimate the gene expression profiles and the parameters are estimated online using Kalman filter. In order to capture the inherent sparsity of the gene networks, LASSO based least square optimzation is performed. The performance of the proposed algorithm is compared with the extended Kalman filter (EKF) algorithm using Mean Square Error (MSE) as the fidelity criterion. The simulations are performed using the synthetic as well as real data and the proposed algorithm is observed to outperform the EKF in the scenarios considered.

Index Terms— Gene regulatory network, particle filter, Kalman filter, parameter estimation, LASSO.

1. INTRODUCTION

Gene regulatory networks model the interactions among the genes and provide a decision rule describing activation and repression of each gene via various proteins. Gene networks can be modeled in multiple ways varying in their degree of sophistication [1]. Static or dynamic models can be used for inferring gene networks based on the type of experiment and data at hand. Various modeling techniques have been reported in literature including, Bayesian networks, factor graphs, Boolean networks, neural networks etc. Currently, a lot of effort is being devoted to introduce improvements in these algorithms so as to enhance our understanding about gene interactions.

Kalman filter has also been employed to model gene regulatory networks [1], [3] but it is applicable only to linear Gaussian models. In order to be able to capture complex gene interactions efficiently, it is imperative to develop algorithms that model that nonlinear interactions among the genes as well. EKF is one such method for estimation and prediction in case of nonlinearity and has been frequently used to perform inferences in gene networks [4]. An EKF-based approach works well in the presence of steady state data and slow dynamics, because it is only an approximation relying on a first-order linearization of the nonlinear model. To cope with nonlinearities, this paper proposes the usage of particle filter techniques which can model the evolving dynamics of a system by catering for any possible nonlinearity. The microarray data is assumed to obey a linear model. To obtain a more accurate and precise picture of gene interactions, a nonlinear model for gene expressions and a discrete time state space system of equations are considered to model possible time variations.

Since a particular gene interacts with a few other genes only, the parameter vector is expected to be sparse. To capture this sparsity, the particle filter is augmented with a *Least Squares Shrinkage Selection Operator* (LASSO) based least squares regression operation. The performance of the aforementioned algorithm is rigorously evaluated for synthetic as well as real biological data sets arising from Drosophila Melanogaster time series gene expression profiles. The results are contrasted with those reported in [4]. It is demonstrated that particle filtering followed by LASSO outperforms the EKF approach for synthetic as well as real data.

2. SYSTEM MODEL

The dynamical gene system is modeled using a standard state space representation. Assuming a system consisting of N genes, the model for the evolution of states at the ith time instant can be expressed as

$$\mathbf{y}_i = g(\mathbf{y}_{i-1}, \mathbf{w}_i) \tag{1}$$

where the function $g(.)$ characterizes the regulatory relationship among various genes. The state vector $\mathbf{y}_i$ represents the gene expression values at a particular time instant i and the noise $\mathbf{w}_i$ impairing the system is assumed to be $i.i.d$ Gaussian such that $w_{i,n} \sim \mathcal{N}(0, \sigma_w^2)$. The microarray data is represented in terms of the variables $\mathbf{z}_i$ which also constitute a

[*]The work of A. Noor and E. Serpedin was supported by NSF Grant 0915444.

set of noisy observations and is given by

$$\mathbf{z}_i = h(\mathbf{y}_i, \mathbf{v}_i) \tag{2}$$

where $\mathbf{v}_i$ is considered $i.i.d$ Gaussian such that $v_{i,n} \sim \mathcal{N}(0, \sigma_v^2)$. In order to capture the inherent nonlinearity relationships existing among genes, a sigmoid squash function is considered which is given by [4]

$$y_{i,n} = \sum_{m=1}^{N} \frac{b_{nm}}{1 + e^{-y_{i-1,m}}} + w_{i,n}$$
$$i = 1, ..., I, \quad n = 1, ..., N \tag{3}$$

where the constant I denotes the total number of time instants in the observed data set and b_{nm} model the regulatory relations among various genes. A discrete linear Gaussian model for the microarray data is considered which can be expressed at the ith time instant as [3], [4]

$$\mathbf{z}_i = \mathbf{y}_i + \mathbf{v}_i. \tag{4}$$

The system model outlined above is complete in the sense that it captures all important features of the gene regulatory network, e.g., nonlinearity, dynamics, and noise in both, microarray data and gene regulation model.

3. METHOD TO INFER GENE REGULATORY NETWORKS

In this section, the methodology proposed to infer the system parameters in (3) is described. Our approach is best illustrated in Fig.1. The algorithm is presented in detail below.

3.1. Particle Filter

Particle filtering is a suboptimal algorithm which uses point masses along with the associated weights to approximate the probability densities [2]. It serves as a natural candidate for making inference in gene regulatory networks since it can be applied to nonlinear state evolution models.

Let $\mathbf{d}_i$ denote the set of all observations up to time i, i.e., $\mathbf{d}_i \triangleq [\mathbf{z}_1, \ldots, \mathbf{z}_i]^T$. Based on the measurements $\mathbf{d}_i$ and past state estimates $\mathbf{y}_{1:i-1}$, our objective is to estimate the current state $\mathbf{y}_i$. This requires the posterior density $p(\mathbf{y}_i|\mathbf{d}_{i-1})$ of the state $\mathbf{y}_i$ as well as the likelihood density $p(\mathbf{z}_i|\mathbf{y}_i)$. Based on the Markov model in (3), the conditional distribution of the state $\mathbf{y}_i$ can be expressed as

$$p(\mathbf{y}_i|\mathbf{y}_{i-1}; b_{nm}) = \frac{1}{(2\pi\sigma_w^2)^{N/2}} \exp\left\{\frac{-||\mathbf{y}_i - \sum_{m=1}^{N} b_{nm} f(\mathbf{y}_{i-1,m})||^2}{2\sigma_w^2}\right\}. \tag{5}$$

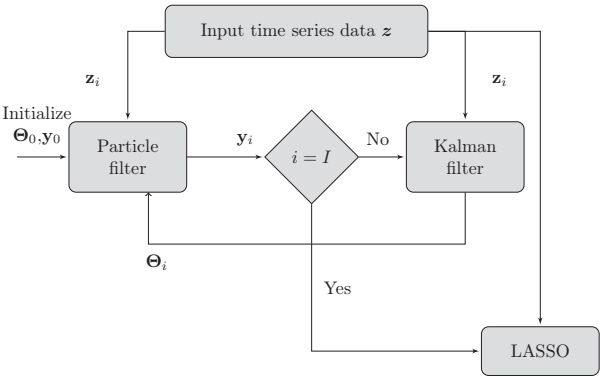

Fig. 1. Block diagram of Gene Network Inference Methodology

and the likelihood density is given by

$$p(\mathbf{z}_i|\mathbf{y}_i) = \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left\{\frac{-||\mathbf{z}_i - \mathbf{y}_i||^2}{2\sigma_v^2}\right\}. \tag{6}$$

The use of these densities to obtain the state estimates is described succinctly in Algorithm 1 (see Subsection 3.4). The process is initiated by drawing K particles $\{\mathbf{y}_0\}_{k=1}^{K}$ for an initial state $\mathbf{y}_0$ from a known prior density $p(\mathbf{y}_0)$ which is assumed Gaussian in this paper. Hence, the posterior density is approximated using these particles and associated weights as

$$p(\mathbf{y}_i|\mathbf{d}_i) = \sum_{k=1}^{K} \xi_i^k \delta(\mathbf{y}_i - \mathbf{y}_i^k).$$

The weights are chosen according to the phenomenon of *importance sampling*, whereby if it is difficult to draw samples from the posterior density, *importance density* is sampled instead [2]. In this paper, the importance density is chosen to be equal to the prior density which provides ease of implementation and hence, the weights update is given by

$$\xi_i^k \propto \xi_{i-1}^k p(\mathbf{z}_i|\mathbf{y}_i^k). \tag{7}$$

In order to tackle the degeneracy phenomenon associated with sequential Monte-Carlo algorithm, resampling is performed [2].

3.2. Kalman Filter

In our framework, particle filter works in conjunction with the Kalman filter, with the former predicting the unknown states $y_{m,i-1}$ and the latter estimating the constant system parameters. It can be observed from the system model described in (3) and (4) that given the states $y_{m,i-1}$, the state space model becomes linear in the unknown parameters. The linearity of this model and Gaussian noise impairment makes Kalman fil-

ter a natural candidate for estimating b_{nm}. Define

$$\Theta \triangleq [b_{11}, \ldots, b_{1N}, b_{21}, \ldots, b_{2N}, \ldots, b_{N1}, \ldots, b_{NN}]^T$$

$$\boldsymbol{f}' \triangleq [f(y_{i-1,1}) \ldots f(y_{i-1,N})] \tag{8}$$

$$\boldsymbol{\Psi}_{f_i} \triangleq \begin{bmatrix} \boldsymbol{f}' & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{f}' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{f}' \end{bmatrix} \tag{9}$$

Then the state and output equations can be expressed as

$$\begin{aligned} \Theta_i &= \Theta_{i-1} + \eta_i \\ \mathbf{z}_i &= \boldsymbol{\Psi}_{f_i} \Theta_i + \mathbf{v}_i \end{aligned} \tag{10}$$

where the parameters Θ_i are assumed to evolve subject to a Gauss-Markov process. The noise η_i denotes the uncertainty in the unknown parameters and its variance is assumed low so that the parameters are almost constant. The parameter set Θ_i is estimated using the standard Kalman filter prediction and update equations.

3.3. Sparse Approximation of Parameters

Kalman filter algorithm provides a solution to a least squares problem, whereby a large number of solutions is possible. LASSO is a solution to a least squares regression problem with an additional L_1 norm penalty on the parameter vector, and hence, provides a unique solution [5]. Therefore, LASSO provides an efficient means of system selection.

After the particle filtering stage delivers the state estimates $y_{i,n}$, they are fed to the LASSO which identifies the system parameters b_{nm} using the estimated states and the observed data. A key rationale behind the proposed LASSO based least squares data fitting is that for a particular gene in question, it is related to only a few other genes and as such, many of the constants b_{nm} signifying the regulation relationship among various genes are 0. LASSO allows to identify only a subset of the system parameters forcing the other irrelevant (or 'weak') interactions to zero. As a result, a parsimonious and efficient description of the gene regulatory network is obtained. For the nth gene, its observations and estimated states can be stacked using (3) and (4) as

$$\begin{bmatrix} z_{n1} \\ z_{n2} \\ \vdots \\ z_{nI} \end{bmatrix} = \begin{bmatrix} f(y_{0,1}) & \ldots & f(y_{0,N}) \\ f(y_{1,1}) & \ldots & \vdots \\ \vdots & \ddots & \\ f(y_{I-1,1}) & & f(y_{I-1,N}) \end{bmatrix} \begin{bmatrix} b_{n1} \\ b_{n2} \\ \vdots \\ b_{nN} \end{bmatrix} + \begin{bmatrix} v_{n1} \\ v_{n2} \\ \vdots \\ v_{nI} \end{bmatrix}$$

which can be compactly expressed as

$$\mathbf{z}_n = \boldsymbol{\Phi} \mathbf{b}_n + \mathbf{v}_n. \tag{11}$$

LASSO operates on this system of equations for the nth gene and produces a parameter vector $\mathbf{b}_n$ by minimizing the criterion:

$$\min_{\mathbf{b}_n} \frac{1}{2} ||\mathbf{z}_n - \boldsymbol{\Phi} \mathbf{b}_n||_2^2 + \lambda ||\mathbf{b}_n||_1. \tag{12}$$

In this paper, the value of λ is considered to be in the range of $[5-10]$. These values were chosen arbitrarily so as to keep the network sparse. The invertibility of the matrix $\boldsymbol{\Phi}^T \boldsymbol{\Phi}$ defined in (11) ensures that the objective function is strictly convex and a globally optimal solution is guaranteed.

3.4. Inference Algorithm

The operation of our algorithm to infer the gene regulatory network is graphically depicted in Fig. 1 and the corresponding pseudocode is summarized in Algorithm 1. In essence, a particle filtering approach to estimate the states coupled with an online Kalman filter based parameter estimation delivers the estimated states to the LASSO operator.

Algorithm 1 `Gene Network Inference`

```
Input time series data set z.
Initialize I, K, θ_0.
for k = 1, ..., K do
    Draw y_0^k ~ p(y_0|z_0).
    Assign weights to the particles using (7).
end for
Calculate the total weight i.e., S = SUM[ξ_i^k]_{k=1}^K
for i = 1, ..., I do
    for k = 1, ..., K do
        Draw y_i^k ~ p(y_i|y_{i-1}, z_i).
        Assign weights to the particles using (7).
    end for
    Calculate the total weight i.e., S = SUM[ξ_i^k]_{k=1}^K.
    Normalize ξ_i = ξ_i / S.
    Estimate parameters b_i from y_i
end for
Resample.
LASSO: Estimate parameters b from y and z using (12).
return
```

4. SIMULATION RESULTS

4.1. Application on Synthetic Data

An eight-gene network is considered for evaluation of the ability to predict the gene expression profiles. The data of length I = 40 is generated using the model given in (3). The variance of the system noise $v_n \sim \mathcal{N}(0, \sigma_v^2)$ is taken to be 10^{-4} and the variance of measurement noise w_n is varied from 10^{-4} to 10^{-1} which is assumed to be known a-priori. For particle filter, number of particles used is $K = 100$ in all simulations. The prediction estimates for both the algorithms, using the inferred system parameter values, b_{nm}, are calculated and compared by using the MSE as the fidelity criterion and the results are shown in Fig. 2. To keep the paper concise, MSE is shown for two representative genes only.

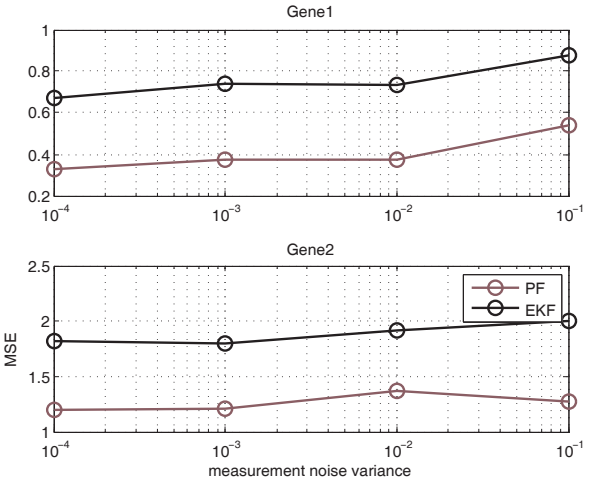

Fig. 2. MSE performance comparison between EKF and PF with LASSO for synthetic data

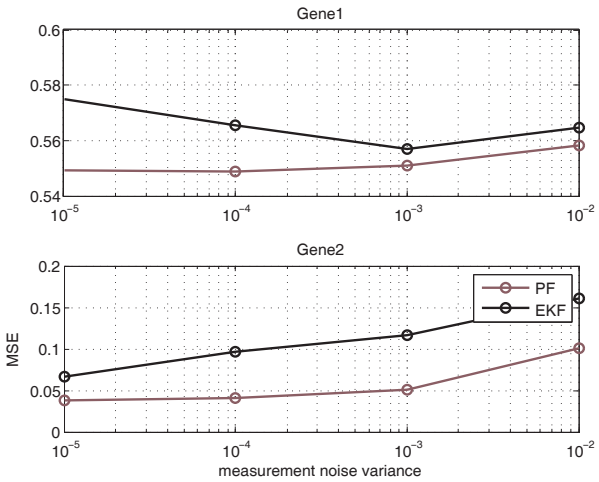

Fig. 3. MSE performance comparison between EKF and PF with LASSO for real data

4.2. Application on the Drosophila Dataset

A time series data set for heat stress response in *drosophila melanogaster* is considered which consists of 36 time points for 530 genes [6]. A data set consisting of ten genes is considered and the evaluation of the algorithm is performed by comparing the predicted values of gene expression profiles by using MSE as the fidelity criterion. The system noise is taken to be $v_n \sim \mathcal{N}(0, 10^{-4})$ and the observation noise $w_n \sim \mathcal{N}(0, \sigma_w^2)$ variance ranges from 10^{-4} to 10^{-1}. It is found that particle filter outperforms the EKF approach consistently for the entire range of observation noise variances as shown in Fig. 3. It can be inferred that the proposed algorithm models the network efficiently and is robust to changes in noise. As in the previous section, a representative figure for two genes is given to keep the paper concise.

5. CONCLUSIONS

This paper considers the modeling and learning of gene regulatory networks using a nonlinear dynamical model which represents a more general scenario. The gene network is modeled using a state space approach and particle filtering is used for state estimation. The parameters regulating the interaction among genes are supplied by an online Kalman filter. Since the parameter vector is frequently sparse, LASSO identifies a subset of these parameters giving only the relevant system coefficients. Extensive performance evaluations demonstrate that this particle filter based approach outperforms the EKF in terms of the MSE criterion. The results are proved using synthetic data as well as real data for Drosophila Melanogaster gene expression time series.

6. REFERENCES

[1] Y. Huang, I. M. Tienda-Luna, and Y. Wang, "Reverse Engineering Gene Regulatory Network," *IEEE Sign. Process. Mag.*, pp. 76-97, Jan 2009

[2] P. M. Djuric, J. H. Kotecha, J. Zhang, Y Huang, T Ghirmai, M. F. Bugallo and J. Miguez, "Particle Filtering" *IEEE Sign. Process. Mag.,* pp. 19-38, Sep 2003.

[3] L. Qian, H. Wang, and E. R. Dougherty, "Inference of Noisy Nonlinear Differential Equation Models for Gene Regulatory Networks using Genetic Programming and Kalman Filtering", *IEEE Trans. Sign. Process.,* vol.56, no.7, pp. 3327-3339, July 2008.

[4] Z. Wang, X. Liu, Y. Liu, J. Liang, and V. Vinciotti, "An Extended Kalman Filtering Approach to Modeling Nonlinear Dynamic Gene Regulatory Networks via Short Gene Expression Time Series", *IEEE Trans. Comp. Bio. and. Bioinf.,* vol.6, no. 3, pp. 410-419, July-Sep 2009.

[5] R. Tibshirani, "Regression Shrinkage and Selection via the LASSO", *J. Royal Statist. Soc B.*, vol. 58, no. 1, pp. 267-288, 1996.

[6] J. C. Costello, et. al. "Gene Networks in Droshophila Melanogaster: Integrating Experimental data to Predict Gene Function," *Genome Biology*, vol. 10, Issue 9, pp. R97.1-97.29, 2009.

[7] M. Gustafsson, M. Hornquist, and A. Lombardi, "Constructing and Analyzing a Large-Scale Gene-to-Gene Regulatory Network - Lasso-Constrained Inference and Biological Validation", *IEEE Trans. Comp. Bio. and. Bioinf.,* Vol. 2, No. 3, pp. 254-261, July-Sep. 2005.