INFERENCE OF GENETIC REGULATORY NETWORKS USING REGULARIZED LIKELIHOOD WITH COVARIANCE ESTIMATION

Ghulam Rasool[1], *Nidhal Bouaynaya*[1], *Hassan M. Fathallah-Shaykh*[2] *and Dan Schonfeld*[3]

[1] Department of Systems Engineering, University of Arkansas at Little Rock
[2] Department of Neurology, University of Alabama at Birmingham
[3] Department of Electrical and Computer Engineering, University of Illinois at Chicago

ABSTRACT

We cast the problem of reverse-engineering the connectivity matrix of genetic regulatory networks from a limited number of measurements as a regularized multivariate regression problem. The regularization term incorporates the prior knowledge of sparsity of genetic regulatory networks. Moreover, the genetic profiles within a measurement are assumed to be correlated with a full covariance structure. The proposed algorithm computes a sparse estimate of the connectivity matrix that accounts for correlated errors using regularized likelihood. We show that the joint estimation of the connectivity and covariance matrices improves the estimation of the network connectivity as compared to the assumption of uncorrelated measurements. Our algorithm has $\ln(\ln(N))$ sampling complexity. We test and validate our approach using synthetically generated networks.

Index Terms— Gene regulatory network; multivariate regression; maximum likelihood estimation; convex optimization.

1. INTRODUCTION

The regulatory processes at work in the cell echo the elaborate network of interactions between the genes, RNAs and proteins. Identifying and understanding these interactions is considered as one of the main challenges in systems biology with potential applications in therapeutic targeting and drug design. The recent advances in high-throughput genomic sequencing technology spurred the reverse-engineering of molecular interactions from collected genomic profiles.

Ordinary differential equations (ODEs) can successfully model the dynamics of genetic profiles [1], with several advantages over graphical methods for genetic network inference. First, ODE approaches yield signed directed graphs, where the sign of an edge indicates if the interaction is stimulative or inhibitive and the absolute value of the interaction reveals the strength of the stimulation or inhibition. Second, ODE inference methods can be applied to both steady-state and time-varying genetic data. In particular, they can be used to predict the behavior of the network at any future time point and under any given condition, such as gene knockout or drug delivery. Most ODE system identification methods assume that the behavior of the regulatory network can be modeled by a system of linear differential equations near a steady-state:

$$\dot{x}_i(t_k) = \sum_{j=1}^{N} a_{ij} x_j(t_k) + b_i u(t_k), \tag{1}$$

where $i = 1, \cdots, N, k = 1, \cdots, M$, N is the number of genes, M is the number of experiments or time points, $x_i(t)$ is the expression of gene i at time t, $\dot{x}_i(t)$ is the rate of change of expression of gene i at time t, a_{ij} represents the influence of gene j on gene i, b_i is the effect of the external perturbation on gene i and $u(t)$ denotes the external perturbation at time t. The goal is to infer the gene interactions $\{a_{ij}\}_{i,j=1}^{N}$, given a certain number of measurements M.

The attempts to solve the problem in (1) rely mainly on linear regression to calculate the model coefficients either at steady-state [2] or assuming a dynamic model [3], [4]. However, the ODE model in Eq. (1) is deterministic and does not take into account the biological stochasticity and measurement noise present in the data. De Hoon *et al.* [5] incorporated an error term in order to statistically determine the sparsity of the connectivity matrix $A = \{a_{ij}\}$. They use the Akaike Information Criterion (AIC) to decide which coefficients should be set to zero.

In this paper, we explicitly add a noise term to the linear ODE model in Eq. (1) in order to take into account biological variability and measurement noise. Unlike previous work, which assumed an uncorrelated noise model [5], we assume that the noise has an unknown correlation structure. The assumption of correlated noise emanates from the fact that separately estimating the interaction coefficients by performing M separate regressions may be inferior to jointly estimating all coefficients accounting for the correlated errors. We then propose a regularized likelihood algorithm, which simultaneously estimates the connectivity and the covariance matrices.

2. THE LINEAR ODE MODEL

We consider the model in Eq. (1) with an additive noise term $\epsilon_i(t)$. Introducing the new variable y_i,

$$y_i(t) = \frac{dx_i}{dt} - b_i u(t), \tag{2}$$

we can write the ODE model in vector form for the N genes as

$$\mathbf{y} = A\mathbf{x} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{y} = [y_1, y_2, \cdots, y_N]^T, \mathbf{x} = [x_1, x_2, \cdots, x_N]^T, \boldsymbol{\epsilon} = [\epsilon_1, \cdots, \epsilon_N]^T$ and $A = \{a_{ij}\}_{i,j=1}^N$. Performing M different experiments , we obtain M measurements and can write the results as

$$Y = AX + E, \tag{4}$$

where $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_M]$, $X = [\mathbf{x}_1, \cdots, \mathbf{x}_M]$ and $E = [\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_M]$. That is, every column of Y, X, and E represents a single experiment and there are M columns representing M experiments. The goal of reverse-engineering the network is to estimate the connectivity matrix A given a number of measurements and in the presence of noise.

3. REGULARIZED LIKELIHOOD WITH COVARIANCE ESTIMATION

We assume that $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_M$ are i.i.d $\mathcal{N}(\mathbf{0}, \Sigma)$. Thus, for any experiment, the covariance matrix of the gene expressions is Σ.

Lemma 1 *The negative log-likelihood function of (A, Σ) can be expressed up to a constant as*

$$-l(A, \Sigma) = Tr[\frac{1}{M}(Y - AX)(Y - AX)^T\Sigma^{-1}] + \ln|\Sigma|, \tag{5}$$

where Tr denotes the trace function and $|\Sigma|$ is the determinant of the matrix Σ.

Proof 1 *We have*

$$p(\boldsymbol{y}_j) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}} \exp -\frac{1}{2}\left[(\boldsymbol{y}_j - A\boldsymbol{x}_j)^T\Sigma^{-1}(\boldsymbol{y}_j - A\boldsymbol{x}_j)\right]. \tag{6}$$

Therefore, the likelihood function of the data is

$$\begin{aligned} p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M) &= \prod_{j=1}^M p(\boldsymbol{y}_j) \\ &= \frac{1}{(2\pi)^{\frac{MN}{2}}|\Sigma|^{\frac{M}{2}}} \exp -\frac{1}{2}\sum_{j=1}^M \left[(\boldsymbol{y}_j - A\boldsymbol{x}_j)^T\Sigma^{-1}(\boldsymbol{y}_j - A\boldsymbol{x}_j)\right] \\ &= \frac{1}{(2\pi)^{\frac{NM}{2}}|\Sigma|^{\frac{M}{2}}} \exp -\frac{1}{2} Tr\left[(Y - AX)(Y - AX)^T\Sigma^{-1}\right] \end{aligned} \tag{7}$$

Taking the logarithm to compute the log-likelihood function, we obtain

$$\begin{aligned} l(A, \Sigma) = &-\frac{MN}{2}\ln(2\pi) - \frac{M}{2}\ln|\Sigma| \\ &-\frac{1}{2} Tr\left[(Y - AX)(Y - AX)^T\Sigma^{-1}\right] \end{aligned} \tag{8}$$

Ignoring the constant term $\left(-\frac{MN}{2}\ln(2\pi)\right)$, which will have no effect in the optimization over A and Σ, and multiplying by $2/M$, we obtain the negative log-likelihood function up to a constant term

$$-l(A, \Sigma) = \frac{1}{M} Tr\left[(Y - AX)(Y - AX)^T\Sigma^{-1}\right] + \ln|\Sigma| \tag{9}$$

Assuming that $M \geq N$, it is easy to derive that the maximum likelihood estimator for A is simply given by $\hat{A} = \hat{A}^{OLS} = YX^T(XX^T)^{-1}$, which ammounts to performing separate ordinary least-squares estimates for each of the measurements, and is independent of the covariance structure. In order to exploit the correlation structure to improve the prediction performance, we impose a constraint on the connectivity matrix A. It is known that genetic regulatory networks are sparse, where each gene is regulated by only few other genes [1]. This prior knowledge, along with the estimation of the correlation structure, will improve the estimation of the network connectivity. To this aim, we consider the regularized negative likelihood function

$$\begin{aligned} f(A, \Omega) = &\mathrm{Tr}\left[\frac{1}{M}(Y - AX)(Y - AX)^T\Omega\right] - \ln|\Omega| \\ &+ \alpha\sum_{i=1}^N\sum_{j=1}^N |a_{i,j}|, \end{aligned} \tag{10}$$

where $\Omega = \Sigma^{-1}$ and $\alpha \geq 0$ is a tuning parameter. The added penalty term is reminiscent of the L_1 norm constraint, which tends to force many entries of the matrix A to be zero; thus achieving a sparse solution. The regularized maximum likelihood estimates of A and Ω are, therefore, given by

$$(\hat{A}, \hat{\Omega}) = \underset{A,\Omega}{\mathrm{argmin}} f(A, \Omega) \tag{11}$$

The optimization problem in (11) is not convex; however, solving for either A or Ω with the other fixed is convex. In particular, the convexity of $f(A)$ for constant Ω follows from the fact that Ω is positive semi-definite. We propose an iterative minimization procedure with respect to A and Ω as outlined below. In our simulations, we use the cvx environment [6] to solve the convex optimization problems involved in the steps of the algorithm.

4. SIMULATION RESULTS

We compare the proposed reguralized likelihood with covariance estimation algorithm to the case where the correlation

For a fixed value of α, set $A^{(0)} = \mathbf{0}$ and $\Omega^{(0)} = I$ (the identity matrix). Set $m = 1$.

Step 1 Compute $\hat{A}^{(m)}$ as the unique solution of the convex optimization problem in (10) for the given $\hat{\Omega}^{(m-1)}$.

Step 2 Compute $\hat{\Omega}^{(m)}$ as the unique positive semi-definite solution of the convex optimization problem in (10) for the $\hat{A}^{(m)}$ computed in step 1.

Step 3 If $\sum_{ij} |\hat{a}_{ij}^{(m)} - \hat{a}_{ij}^{(m-1)}| < \eta$, then stop, otherwise, increment m and go to step 1.

structure between the genes is ignored, i.e., $\Sigma = \sigma^2 I$. We use Eq. (4) to generate synthetic networks with varying size, number of measurements and correlation structure. We introduce sparsity in the network by setting $A_{ij} \neq 0$ for $N^2 \times d$ elements, where N represents the number of genes in the network and d is a percentage number. In our simulations, we use $d = 5\%$. The non-zero elements of the matrix A are drawn from a standard normal distribution with zero mean and unit variance, i.e., $a_{ij} \in \mathcal{N}(0, 1)$, for all $a_{ij} \neq 0$. The covariance matrix is generated using the following model $\Sigma_{ij} = \rho^{|i-j|}$, where $\rho = 0.9$ is a fixed correlation coefficient. To assess the efficiency of the algorithm, we use the following performance measure proposed in [3]

$$E = \sum_{i=1}^{n} \sum_{j=1}^{n} e_{ij}, \quad \text{with}$$
$$e_{ij} = \begin{cases} 1, & \text{if} \quad |A_{R,ij} - A_{T,ij}| > \delta, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where A_R and A_T denote, respectively, the estimated and true connectivity matrices, and δ is a fixed threshold. In our simulations, we set $\delta = \frac{1}{2} \min_{i,j} \{|a_{ij}| \neq 0\}$.

In the current application, the number of measurements is smaller than the number of genes, i.e., $N > M$. Nonetheless, we rely on the proposed approach for the experimental study. We first investigate the influence of the sparsity coefficient α on the estimation error. Figure 2 shows the number of errors, E, as a function of α for $N = 20, 30, 40$-gene networks. Observe that the estimation error decays rapidly from $\alpha = 0$ to $\alpha \neq 0$. In fact, when $\alpha = 0$, there are no constraints on the network connectivity, and the ML estimate of A is independent of the correlation structure. When α is very large, the estimate of the connectivity matrix is basically the L_1 norm estimator, i.e., the zero matrix. We found that values of α in the range 0.1 to 1 provide a good balance between likelihood and sparsity considerations. In our simulations, we set $\alpha = 0.5$. Figure 2 also shows that, for a given value of α and fixed number of measured, the error increases as the number of genes increases. This result is further illustrated in Fig. 3, where it is seen that the estimation error decreases rapidly as the number of emasurements increases.

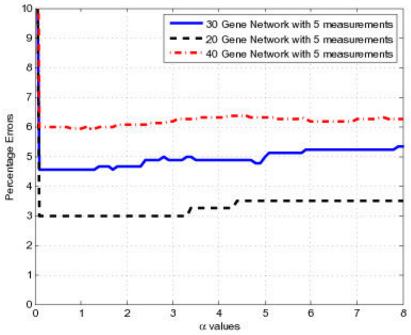


Fig. 2. The effect of the regularization parameter α on the estimation accuracy of the connectivty matrix A.

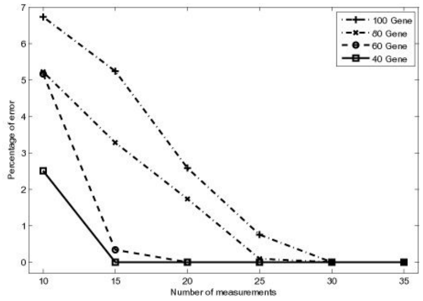
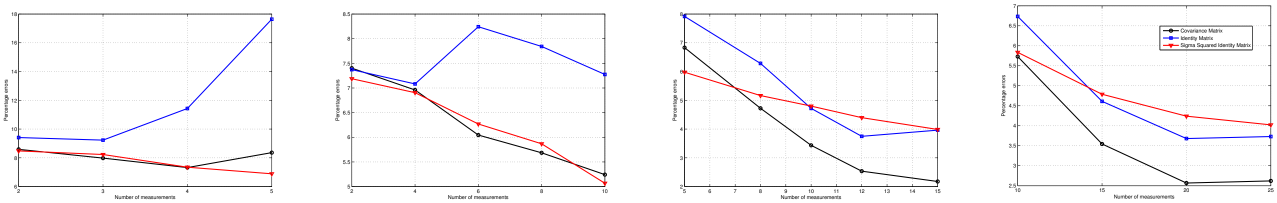


Fig. 3. Percentage error versus number of measurements for different network sizes.

We further evaluate the number of measurements necessary to identify the network connectivity with 99% confidence as a function of the network size N. For a given network size, the smallest number of measurements required to recover the connectivity of the network with an error below 1% is called the critical number of measurements, M_c. To obtain statistically meanigful claims, we perform Monte Carloe simulations for 100 realizations of the network. We found that the number of critical measurements increases linearly with $\ln(\ln(N))$. The least-squares fit curve, of the form $M_c = a + b \ln(\ln(N))$, is displayed in Fig. 4. In particular, for a 1000-gene network, the proposed algorithm requires about 56 measurements to correctly identify the network topology. This critical number is significantly smaller than requiring 1000 measurements in the “brute-force approach”, or the 90 measurements required when using the singular value decomposition method [3].

In order to outline the importance of taking into account the correlation structure of the measurements, we compare the performance of the joint covariance estimation with the case where the gene expressions are assumed to be uncorrelated, namely $\Sigma = I$ and $\Sigma = \sigma^2 I$, where σ^2 is estimated using the ML approach. Figure 1 shows the percentage error versus the number of measurements for the three cases for four different network sizes $N = 10, 20, 30$ and 40. It is interesting to ob-



(a)$N = 10$ (b) $N = 20$ (c) $N = 30$ (d)$N = 40$

Fig. 1. Performance comparison of the regularized ML estimation with and without estimation of the covariance structure for different network sizes: black: covariance structure estimation; blue: $\Sigma = I$; red: $\Sigma = \sigma^2 I$.

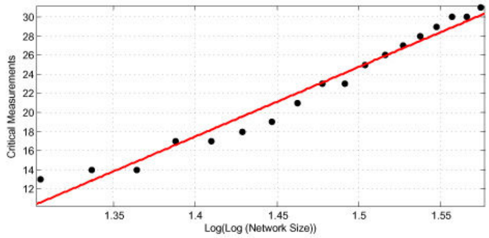


Fig. 4. Critical number of measurements required to recover the network connectivity with 99% confidence. Circles: numerical data; Line: least-squares fit of the form $M_c = a + b\ln(\ln(N))$.

serve that for small networks, the ML estimation of the power of the uncorrelated noise performs as good as estimating the full correlation structure. This is partly due to the fact that, in small networks, the number of measurements is also small, and hence the estimation of the full correlation structure may not be statistically meaningfull. However, for larger networks ($N \geq 30$), joint estimation of the connectivity and covariance matrices yields significantly smaller errors than the connectivity estimation considering uncorrelated measurements.

5. CONCLUSION

We casted the reverse-engineering problem of the network connectivity as a regularized multivariate regression problem with a full covariance structure. Our simulations show that the regularized likelihood with covariance estimation method outperforms both (non regularized) likelihood estimation and uncorrelated regression. We also explored the effect of the number of measurements on the estimation error and evaluated that the critical number of measurements required to recover the entire connectivity matrix within 99% confidence scales as $\ln(\ln(N))$, where N is the number of genes. We, therefore, expect that the proposed reverse-engineering method will be useful in reconstructing gene networks when more experimental data becomes available.

6. ACKNOWLEDGEMENT

This project is supported by Award Number R01GM096191 from the National Institute Of General Medical Sciences (NIH/NIGMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of General Medical Sciences or the National Institutes of Health.

This work is also supported in part by the National Science Foundation under Grant CRI CNS-0855248, Grant EPS-0701890, Grant EPS-0918970, Grant MRI CNS-0619069, and OISE-0729792.

7. REFERENCES

[1] F d'Alché Buc, *Biological Networks*, vol. 3, chapter Inference of Biological Regulatory Networks: Machine Leariningg Approaches, pp. 41–82, World Scientific, 2007.

[2] T S. Gardner, D di Bernardo, D Lorenz, and J J Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science*, vol. 301, no. 5629, pp. 102–105, July 2003.

[3] M K S Yeung, J Tegner, and J J Collins, "Reverse engineering gene networks using singular value decomposition and robust regression," *PNAS*, vol. 99, no. 9, pp. 6163–6168, April 2002.

[4] M Bansal, G D Gatta, and D di Bernardo, "Inference of gene regulatory networks and compound mode of action from time course gene expression profiles," *Bioinformatics*, vol. 22, no. 7, pp. 815–822, 2006.

[5] M J De Hoon, S Imoto, K Kobayashi, N Ogasawara, and S Miyano, "Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations," in *Pacific Symposium on Biocomputing*, 2003, pp. 17–28.

[6] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," Apr. 2011.