Inference of Genetic Regulatory Networks with Unknown Covariance Structure

Belhassen Bayar and Nidhal Bouaynaya
Department of Electrical and Computer Engineering
Rowan University
Glassboro, NJ

Roman Shterenberg
Department of Mathematics
University of Alabama at Birmingham
Birmingham, AL

***Abstract*—The major challenge in reverse-engineering genetic regulatory networks is the small number of (time) measurements or experiments compared to the number of genes, which makes the system under-determined and hence unidentifiable. The only way to overcome the identifiability problem is to incorporate prior knowledge about the system. It is often assumed that genetic networks are sparse. In addition, if the measurements, in each experiment, present an unknown correlation structure, then the estimation problem becomes even more challenging. Estimating the covariance structure will improve the estimation of the network connectivity but will also make the estimation of the already under-determined problem even more challenging. In this paper, we formulate reverse-engineering genetic networks as a multiple linear regression problem. We show that, if the number of experiments is smaller than the number of genes and if the measurements present an unknown covariance structure, then the likelihood function diverges, making the maximum likelihood estimator senseless. We subsequently propose a normalized likelihood function that guarantees convergence while keeping the form of the Gaussian distribution. The optimal connectivity matrix is approximated as the solution of a convex optimization problem. Our simulation results show that the proposed maximum normalized-likelihood estimator outperforms the classical regularized maximum likelihood estimator, which assumes a known covariance structure.**

***Index Terms*—Gene regulatory networks; Maximum likelihood estimation; Under-determined systems.**

I. Introduction

Inferring gene regulatory networks from high-throughput data is an important subject in computational systems biology because it renders possible the study of gene-to-gene interactions, the identification of functional modules and the prediction of the behavior of the system under different conditions such as perturbations. A wealth of approaches were suggested to infer genetic regulatory networks including Boolean networks, Bayesian networks, graphical-based model approaches, information-theoretic-based methods and differential equations modeling. Among these approaches, only the class of differential equations presents a continuous model that is able to quantify the direction, strength and sign of the interactions. Identifying whether a genetic interaction is stimulatory or inhibitory is of great importance for understanding the dynamics of the genetic networks and designing appropriate biological experiments for hypothesis validation.

The main difficulty in the problem of inference of genetic regulatory networks, using differential equations (or other methods), is the large number of genes p compared to the number of experiments n. This large p, small n issue makes the problem unidentifiable, i.e., there are many network structures that fit the given data, and no unique solution exists. The only known way to overcome an under-determined problem is to introduce prior knowledge about the system. In genetic networks, it is often assumed that the network is sparse [1]. This assumption is biologically relevant (at least for large networks). It is acknowledged that a gene usually interacts with only a small number of genes in the network.

Since genetic expression data is very noisy (due to measurement uncertainties and inherent stochasticity of biological signals), stochastic ODEs are better suited to model genetic networks, where an error term is added to the deterministic ODE model [1], [2]. We model the dynamics of genetic networks using a system of linear differential equations near a steady-state [1],

$$Y = AX + E, \tag{1}$$

where $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_n]$, $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ and $E = [\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n]$ are $p \times n$ matrices ($p > n$), where every column of Y, X, and E represents a single experiment and there are n columns representing n experiments. $A \in \mathbb{R}^{p \times p}$ is the connectivity matrix of the network. Positive and negative values of A are interpreted as stimulatory and inhibitory interactions, respectively. The column $\boldsymbol{x}_k \in \mathbb{R}^p$, $k = 1, \cdots, n$ contains the expressions of the p genes in the k^{th} experiment. $Y = \dot{X} - U$, where U may be some known control or basal production rate. Model (1) states that the production rate of molecule i is a linear combination of the expressions of all other molecules in the network plus some known basal production rate plus an error term [1].

In order to take into account the noise statistics, a maximum likelihood (ML) approach can be adopted [1], [2]. In [2], the noise is assumed to be uncorrelated. Though Rasool *et al.* [1] proposed a method to estimate the matrix connectivity of the network while accounting for possible correlations in the measurements, their approach, as the authors state, applies only to over-determined systems and cannot be used for under-determined system.

Problem Formulation: Why the maximum likelihood method with covariance estimation is senseless for under-determined systems

Assume that $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_M$ are i.i.d $\mathcal{N}(\mathbf{0}, \Sigma)$. Then, the negative log-likelihood function of (A, Σ) can be expressed up to a constant as [1]

$$-l(A, \Sigma) = \mathrm{Tr}[\frac{1}{M}(Y - AX)(Y - AX)^T \Sigma^{-1}] + \ln|\Sigma|, \quad (2)$$

where Tr denotes the trace function and $|\Sigma|$ is the determinant of the matrix Σ. Because the system is under-determined, there exist solutions satisfying $Y = AX$ and Σ infinitely small. For these solutions, the negative log-likelihood in (2) tends to $-\infty$. Hence, the likelihood, as a function of the two variables (A, Σ), diverges. Observe that the likelihood converges if the covariance matrix Σ is known (e.g., proportional to the Identity for uncorrelated measurements as in [2]) or if the system is over-determined (in this case, there exists no solution that satisfies $Y = AX$).

Assuming uncorrelated observations amounts to separately estimating the regression coefficients (A) by performing n separate regressions. This is inferior than jointly estimating all coefficients when taking into account the correlation in the observations or measurements. Rothman *et al.* [3] proposed a regularized algorithm that simulatenously infers the regression coefficient matrix A and the inverse error covariance, $\Omega = \Sigma^{-1}$, by imposing sparsity constraints on Ω. The l_1-norm penalty on Ω ensures the convergence of the regularized likelihood because it excludes exact solutions, for which the covariance is infinitely small or equivalently the inverse covariance is infinitely large. However, in many applications, the assumption of a sparse inverse covariance matrix may not be reasonable or have any physical justification. In particular, in the genetic regulatory network problem, there is no evidence for such an assumption. Moreover, the solution to the regularized problem in [3] relies on an iterative procedure that finds the maximum over A then over Ω. That is because the problem is convex in each variable, A and Ω, but not convex in the pair (A, Ω). This iterative procedure is not guaranteed to converge and if it does converge, then it may not reach the optimal solution. Hence, the open question remains: "How can we perform maximum likelihood with covariance estimation for under-determined systems?"

This paper addresses this question, namely the problem of ML estimation with unknown covariance in under-determined systems. We present a normalization of the likelihood function that guarantees convergence while still keeping the form of the Gaussian distribution.

The proofs of the mathematical claims will be presented in the extended journal version.

II. The Normalized Likelihood

We define the normalized-likelihood of the under-determined ($n < p$) multiple regression model in (1), under the Gaussian assumption, as

Definition 1.

$$L_N(A, \Omega) = \frac{|(Y - AX)^T \Omega (Y - AX)|^{\frac{n}{2}}}{(2\pi)^{\frac{np}{2}}} \exp -\frac{1}{2} \mathrm{Tr}[(Y - AX)^T \Omega (Y - AX)], \quad (3)$$

where $|\cdot|$ is the matrix determinant operator.

Obviously, one can propose many possible normalizations of the Gaussian likelihood as a function of the pair (A, Ω). Our particular "choice" in Definition 1 is motivated by finding a function that guarantees the convergence of the likelihood while keeping the form of the Gaussian density. The pair (A, Ω) can then be computed to maximize the normalized-likelihood, L_N, i.e.,

$$(A^*, \Omega^*) = \arg\max_{A,\Omega} L_N(A, \Omega), \quad (4)$$

It can be shown that the solution to (4) is given by

$$(Y - A^*X)^T \Omega^* (Y - A^*X) = nI, \quad (5)$$

where I denotes the $n \times n$ Identity matrix.

There are (infinitely) many pairs (A, Ω) that satisfy the optimality condition in (4). In order to obtain a unique solution, we need to further constrain the problem. Among all possible solutions of (5), we find those that minimize the regularized least-square error $\|Y - AX\|_F^2 + \lambda \|\Omega\|_F^2$, where λ is a regularization parameter and $\|\cdot\|_F$ denotes the Frobenius norm. Thus, the optimization problem becomes

$$\begin{cases} \min\limits_{(A,\Omega)} \|Y - AX\|_F^2 + \lambda \|\Omega\|_F^2 \\ \text{s.t.} \quad (Y - AX)^T \Omega (Y - AX) = nI. \end{cases} \quad (6)$$

To solve (6), we use the polar decomposition of matrices.

Definition 2. *The polar decomposition of a matrix $B \in \mathbb{C}^{p \times n}$ is given by*

$$B = U|B|, \quad (7)$$

where $|B| = (B^T B)^{1/2}$, $(\cdot)^{1/2}$ is the principal square root operator and $U : \mathbb{C}^p \longrightarrow$ Range(B) is a $\mathbb{C}^{p \times n}$ isometry such that $U^T U = I$.

Replacing the matrix $(Y - AX)$ by its polar decomposition in (5), we obtain an analytical expression of Ω in terms of A,

$$\Omega_A = n\, U[(Y - AX)^T (Y - AX)]^{-1} U^T, \quad (8)$$

where U is the isometry of $(Y - AX)$. The optimization problem in (6) becomes then equivalent to

$$\begin{cases} \min\limits_{S} \mathrm{Tr}(S) + n^2 \mathrm{Tr}(S^{-2}) \\ \text{s.t.} \quad S = (Y - AX)^T (Y - AX) \end{cases} \quad (9)$$

The objective function in (9) is convex and depends only on one variable S. However, the problem is still not convex because the equality in the constraint is quadratic [4]. The regression coefficient matrix A is sparse, therefore its l_1 norm is upper bounded [5]. We use the sparsity of A and

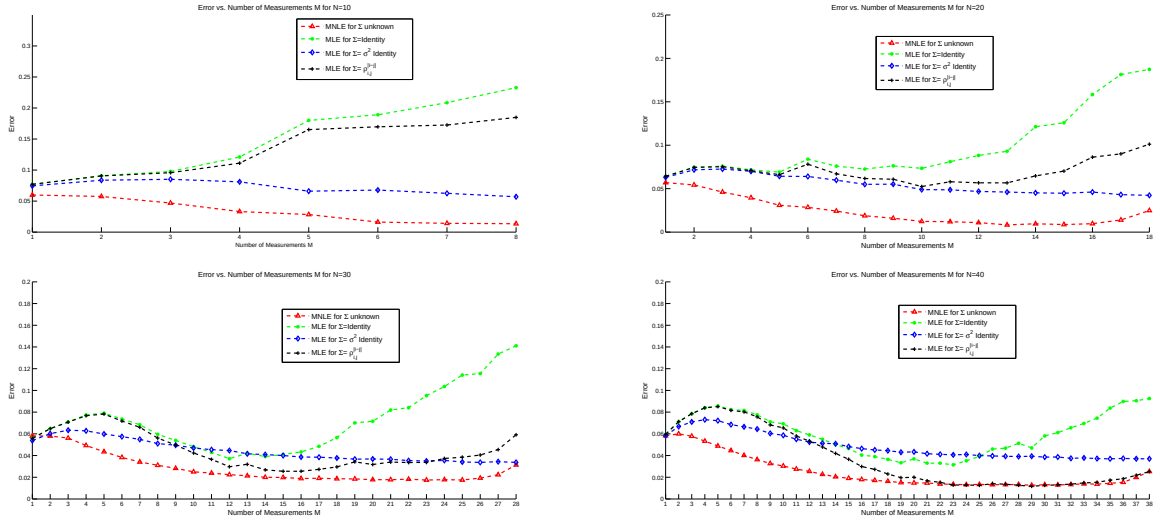

Fig. 1. Performance comparison of the proposed maximum normalized-likelihood (MNL) algorithm with the regularized ML estimation for different network sizes %95 sparse: Red: MNLE with unknown covariance Σ; Green: MLE for $\Sigma = I$; Blue: MLE for $\Sigma = \sigma^2 I$; Black: MLE for $\Sigma = \rho^{|i-j|}$. (a) $p = 10$; (b) $p = 20$; (c) $p = 30$; (d) $p = 40$.

approximate the problem in (9) by a convex optimization problem. If $\|A\|_1 \leq \epsilon$, then the solution to the optimization problem in (9) can be approximated by the solution to the following convex optimization problem

$$\begin{cases} \min\limits_{S} \operatorname{Tr}(S) + n^2 \operatorname{Tr}(S^{-2}) \\ \text{s.t.} \quad S \in \Lambda = \{S \in \mathbb{S}_{n,n} \,|\, \|S - Y^T Y\| \leq \epsilon c\} \end{cases} \tag{10}$$

where $\mathbb{S}_{n,n}$ is the set of $n \times n$ symmetric positive definite matrices and c is a small term which depends on X and Y. Let S^* be the unique global solution of the convex optimization problem in (10). The optimal connectivity matrix, A^*, satisfies, $S^* = (Y - A^*X)^T(Y - A^*X)$. Since there are still many possible solutions to the optimal connectivity matrix, we propose to find the one with minimum l_1 norm,

$$\begin{cases} \min\limits_{A} \|A\|_1 \\ \text{s.t.} \quad AX = Y - U(S^*)^{1/2}, \end{cases} \tag{11}$$

where the equality constraint in Eq. (11) comes from the polar decomposition of $(Y - AX)$. Since A is sparse, we can approximate the solution constructed with U, the isometry of $(Y - AX)$, by a solution constructed by V, the isometry of Y. This approximation reduces the set over which we minimize but leads to an affine equality constraint and hence a convex problem. Thus, to find $\hat{A}$ we solve

$$\begin{cases} \min\limits_{A} \|A\|_1 \\ \text{s.t.} \quad AX = Y - V(S^*)^{1/2}, \end{cases} \tag{12}$$

MNLE algorithm: The MNLE algorithm is summarized below.

Input: The matrices $X \in \mathbb{R}^{p \times n}$ and $Y \in \mathbb{R}^{p \times n}$ ($p > n$) satisfying the under-determined regression model $Y = AX + E$ with an unknown covariance matrix.

Step 1 Solve the convex optimization problem in (10) (using for instance the yalmip package in MATLAB [6]). The solution is a s.p.d. matrix $S^* \in \mathbb{R}^{n \times n}$

Step 2 Given S^*, the optimal connectivity matrix A^* is obtained as the solution to the convex optimization problem in (12) (using for instance, the *cvx* package [7]).

III. Simulation Results

We compare the proposed MNLE algorithm to the regularized MLE algorithm in [1], where the lasso penalty is imposed on the connectivity matrix A. To this aim, we generate synthetic genomic networks with varying size p, number of measurements n and correlated structure Σ. We use the same covariance matrix in [1] where $\Sigma_{i,j} = \rho^{|i-j|}$ and $\rho = 0.9$ is a fixed correlation structure. We use two sparse models of the connectivity matrix, $\|A\|_0 = 0.05p^2$ and $\|A\|_0 = 0.2p^2$, where $\|\cdot\|_0$ is the number of non-zero elements. The entries of the matrix A are drawn from a standard normal distribution with zero-mean and unit variance, i.e., $a_{i,j} \in \mathcal{N}(0,1)$. The performance of the algorithm is assessed through the following measure suggested in [8]

$$E = \sum_{i=1}^{n} \sum_{i=j}^{n} e_{i,j} \quad \text{with} \quad e_{i,j} = \begin{cases} 1, & \text{if } |a_{ij} - \hat{a}_{ij}| > \delta \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where a_{ij} and $\hat{a}_{ij}$ denote, respectively, the true and estimated connectivity entries. δ is a fixed threshold set to $\delta = \frac{1}{2} \min\limits_{i,j} |a_{i,j}| \neq 0$. The percentage error is equal to E/n^2.

Figure 1 shows the percentage error versus the number of measurements n for $p = 10, 20, 30, 40$-gene networks and

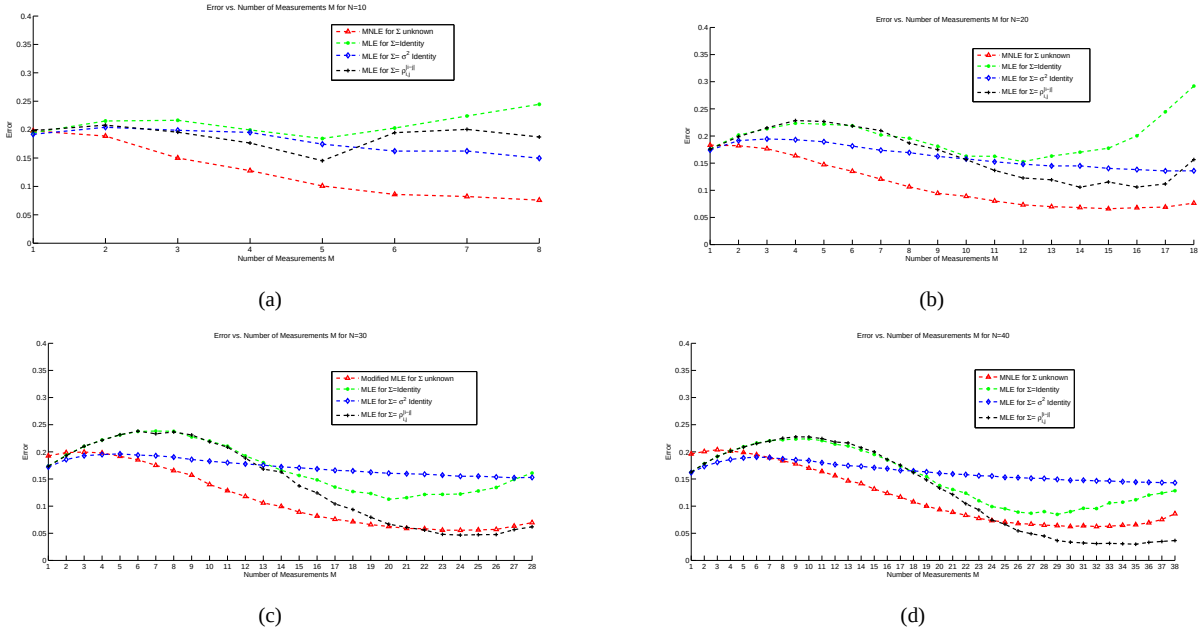

Fig. 2. Performance comparison of the MNLE with the regularized ML estimation for different network sizes $\%80$ sparse: Red: MNLE for Σ unknown; Green: MLE for $\Sigma = I$; Blue: MLE for $\Sigma = \sigma^2 I$; Black: MLE for $\Sigma = \rho^{|i-j|}$. (a) $p = 10$; (b) $p = 20$; (c) $p = 30$; (d) $p = 40$.

$\|A\|_0 = 0.05p^2$. The proposed MNLE algorithm (in red) outperforms the regularized maximum likelihood estimator (ML) with known covariance matrix, where $\Sigma = I, \sigma^2 I, \rho^{|i-j|}$ [1]. 100 Monte Carlo simulations were performed for each curve. Observe that the percentage error of the MNLE is always less than $\%6$ and stabilizes under 1%. Figure 2 shows the same simulations but with a degree of sparsity 80% for the connectivity matrix A. The performance may seem to deteriorate for denser matrices; this may be due to two reasons: First, the MNLE algorithm is built to find the sparsest matrices. Second, the number of errors increases with the number of non-zero elements in the matrix. The proposed MNLE algorithm still outperforms the regularized ML estimator with known covariance matrix.

IV. Conclusion

The maximum likelihood estimator of under-determined Gaussian systems with unknown covariance is senseless. Nonetheless, in many applications, the observations or measurements present an unknown correlation structure and the system is under-determined because of the difficulty or cost of measurements. This is for instance the case in genetic regulatory networks, where the number of genes is much larger than the number of time point measurements, and where gene expression measurements (at each time point) present an unknown correlation structure. For such applications, the maximum likelihood estimator with unknown covariance diverges.

In this paper, we proposed a new maximum normalized-likelihood estimator (MNLE), that guarantees the convergence of the likelihood and keeps its Gaussian form. We show that the optimal estimator can be approximated as the solution of a convex optimization problem. Our simulation results show that the proposed MNLE algorithm outperforms the regularized maximum likelihood estimator with known covariance structure.

Acknowledgment

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.

References

[1] G. Rasool, N. Bouaynaya, H. M. Fathallah-Shaykh, and D. Schonfeld, "Inference of genetic regulatory networks using regularized likelihood with covariance estimation," in *The 2012 IEEE International Workshop on Genomic Signal Processing and Statistics*, December 2012.

[2] M. D. Hoon, S. Imoto, and S. Miyano, "Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations," *Pacific Symposium on Biocomputing*, pp. 17–28, 2003.

[3] A. J. Rothman, E. Levina, and J. Zhu, "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*, 2010.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[5] M. Davenport, M. Duarte, Y. Eldar, and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012, ch. Introduction to compressed sensing.

[6] J. Löfberg, "YALMIP : A toolbox for modeling and optimization in MATLAB," in *CCA/ISIC/CACSD*, Sep. 2004. [Online]. Available: http://control.ee.ethz.ch/index.cgi?action=details;id=2088;page=publications

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

[8] M. K. S. Yeung, J. Tegner, and J. J. Collins, "Reverse engineering gene networks using singular value decomposition and robust regression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 6163–6168, April 2002.