SMURC: High-Dimension Small-Sample Multivariate Regression with Covariance Estimation

Belhassen Bayar, Nidhal Bouaynaya*, and Roman Shterenberg

***Abstract*—We consider a high dimension low sample-size multivariate regression problem that accounts for correlation of the response variables. The system is under-determined as there are more parameters than samples. We show that the maximum likelihood approach with covariance estimation is senseless because the likelihood diverges. We subsequently propose a normalization of the likelihood function that guarantees convergence. We call this method *SMURC*: Small-sample MUltivariate Regression with Covariance estimation. We derive an optimization problem and its convex approximation to compute SMURC. Simulation results show that the proposed algorithm outperforms the regularized likelihood estimator with known covariance matrix. We also apply SMURC to the inference of the wing-muscle gene network of the *Drosophila melanogaster* (fruit fly).**

***Index Terms*—High dimension low sample size; Multivariate Regression; Maximum Likelihood; Gene Regulatory Network.**

I. Introduction

Many engineering problems are formulated as an inverse problem. Examples in signal processing include source estimation of electroencephalographic (EEG) and magnetoencephalographic (MEG) data and inference or reverse-engineering of genetic regulatory networks from high-throughput gene expression data. These problems are sometimes referred to as *ill-posed* or *ill-defined* because the inverse problem has no unique solution, and there are infinitely many solutions that are equally compatible with the data. For instance, in EEG and MEG source estimation problems, if the source distribution contains more independent parameters than there is independent information in the recorded data, then the sources spatial distribution cannot be estimated. In genomics, the inference of genetic regulatory networks also suffers from the limited number of measurements available to unambiguously estimate the network connectivity. This problem, known as the "large p small n" problem, poses a challenge in estimation due to the identifiability problem, where a large class of solutions is consistent with the measurements and no unique solution exists.

The approaches proposed in the literature to tackle inverse problems can be classified into three groups: (1) the statistical approach, which finds the most likely solution that fits the data and any additional constraints that may be imposed; (2) the minimum norm approach, which finds a solution that is compatible with the data and satisfies additional constraints, e.g., on the amplitudes or covariances of the parameters; (3) the resolution optimization methods, which estimate the parameters as independently as possible from each other. It has been shown in [1] that all these approaches result in the same solution given the same a priori information. Moreover, if no a priori information is available, all three methods are equivalent to the classical minimum norm solution [1].

Let us consider the (under-determined) multivariate regression problem, which generalizes the classical regression problem of one response on p predictors to regressing q responses on p predictors. This model has various applications including genomics [2], neurology [3], imaging [3] and econometrics. Let $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$ denote the predictors, $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iq})$ denote the responses, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{iq})$ the errors for the i^{th} sample. The multivariate regression model is given by

$$\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, n, \tag{1}$$

where $\boldsymbol{A}$ is a $q \times p$ regression matrix and n is the sample size. We make the standard assumption that $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n$ are i.i.d Gaussian with zero mean and covariance matrix $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. The model in (1) can be expressed in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}, \tag{2}$$

where $\boldsymbol{Y}$ is the $q \times n$ response matrix with its i^{th} column $\boldsymbol{y}_i$, $\boldsymbol{X}$ is the $p \times n$ predictor matrix with its i^{th} column $\boldsymbol{x}_i$ and $\boldsymbol{E}$ is the random error matrix. $\boldsymbol{X}$ is assumed to be full-rank. The system is under-determined when there are more parameters than samples, i,e, $q > p > n$.

The negative log-likelihood function of $(\boldsymbol{A}, \boldsymbol{\Omega})$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, can be expressed up to a constant as,

$$g(\boldsymbol{A}, \boldsymbol{\Omega}) = \text{tr}\left[\frac{1}{n}(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})^t \boldsymbol{\Omega} (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})\right] - \log|\boldsymbol{\Omega}|, \tag{3}$$

where tr denotes the trace operator. If $p \leq n$ (complete or over-determined system), the maximum likelihood estimator for $\boldsymbol{A}$ is simply given by $\hat{\boldsymbol{A}}^{\text{OLS}} = \boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$, which is independent of $\boldsymbol{\Omega}$ and amounts to performing q separate ordinary least-squares.

The multivariate regression problem becomes particularly challenging when the system is under-determined as it requires the estimation of pq parameters from $nq < qp$ predictors or $n < p$. Different approaches were proposed to reduce the number of parameters by minimizing (3) under various constraints on the regression matrix $\boldsymbol{A}$. Reduced-rank approaches restrict the rank of the estimated matrix of regression coefficients, $\text{rank}(\boldsymbol{A}) \leq r \leq \min(p, q)$ [4]. The rank can also be reduced by

B. Bayar and N. Bouaynaya (corresponding author) are with the Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ. E-mails: belhassen.bayar@gmail.com; bouaynaya@rowan.edu

R. Shterenberg is with the Department of Mathematics, University of Alabama at Birmingham, AL. E-mail:shterenb@math.uab.edu

imposing a sparsity constraint on the singular values of $\boldsymbol{A}$ [5]. Sparsity can also be imposed to identify the main predictors [2], where a combined constraint function that includes l_1 and l_2 regularization, is used [6]. The l_1 constraint introduces sparsity in the entries of $\boldsymbol{A}$ and the l_2 regularization identifies irrelevant predictors (for all q responses) by introducing zeros for all entries in some rows of $\boldsymbol{A}$. However, all of these approaches do not account for correlated responses.

Exploiting the correlation in the response variables improves the prediction performance. For under-determined problems, however, the maximum likelihood (ML) approach with covariance estimation is senseless because there exist solutions satisfying $\boldsymbol{Y} = \boldsymbol{AX}$ and $\boldsymbol{\Sigma}$ infinitely small. For these solutions, the negative log-likelihood in (3) tends to $-\infty$. Hence, the likelihood, as a function of the two variables $(\boldsymbol{A}, \boldsymbol{\Omega})$, diverges. Observe that the likelihood converges if the covariance matrix $\boldsymbol{\Sigma}$ is known (e.g., proportional to the Identity for uncorrelated measurements) or if the system is over-determined (in this case, there exists no solution that satisfies $\boldsymbol{Y} = \boldsymbol{AX}$).

Rothman *et al.* [7] proposed a regularized algorithm that simultaneously infers the regression coefficient matrix $\boldsymbol{A}$ and the inverse error covariance, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, by imposing sparsity constraints on $\boldsymbol{\Omega}$. The l_1-norm penalty on $\boldsymbol{\Omega}$ 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 [7] relies on an iterative procedure that finds the maximum over $\boldsymbol{A}$ then over $\boldsymbol{\Omega}$. That is because the problem is convex in each variable, $\boldsymbol{A}$ and $\boldsymbol{\Omega}$, but not convex in the pair $(\boldsymbol{A}, \boldsymbol{\Omega})$. This iterative procedure is not guaranteed to converge and if it does converge, then it may not reach the optimal solution. Additionally, the authors observed that this algorithm may take many iterations to converge for high-dimensional data. Subsequently, they proposed an approximate MRCE approach that prematurely terminates the iterative optimization procedure after two iterations.

Recently, Zhang *et al.* [8] proposed the sparse Conditional Gaussian Graphical Model (sCGGM). CGGM formulates the inference problem as a joint probabilistic graphical model. sCGGM minimizes the negative log-likelihood of the data with l_1 penalties on the autocorrelation and cross-correlation precision matrices [8]. The main advantage of CGGM over MRCE is that CGGM leads to a convex problem, whereas the MRCE estimation problem is only bi-convex, not jointly convex. However, as acknowledged by the authors, CGGM and MRCE are so similar that "MRCE was mistakenly called a sparse CGGM" [8]. In essence, both algorithms solve an under-determined linear regression problem by maximizing the Gaussian likelihood subject to sparse constraints on the correlation structure. 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 exponential form of the distribution.

In this paper, scalars are denoted by lower case letters, e.g., n, m; vectors are denoted by bold lower case letters, e.g., $\boldsymbol{x}, \boldsymbol{y}$; and matrices are referred to by bold upper case letters, e.g., $\boldsymbol{A}, \boldsymbol{X}$. $\boldsymbol{I}$ denotes the identity matrix. $\boldsymbol{x}_i$ denotes the i^{th} element of vector $\boldsymbol{x}$ and a_{ij} is the $(i, j)^{th}$ entry of matrix $\boldsymbol{A}$. Throughout the paper, we provide references to known results and limit the presentation of proofs to new contributions.

II. The Normalized-Likelihood

We propose to weight the likelihood function by the "energy" of the error, in order to guarantee the convergence of the energy-weighted likelihood function, while still keeping the exponential form of the density. Specifically, we define the normalized-likelihood of the under-determined ($p > n$) multiple regression model in (2), under the Gaussian assumption, as

Definition 1.

$$L_N(\boldsymbol{A}, \boldsymbol{\Omega}) = \frac{|(\boldsymbol{Y} - \boldsymbol{AX})^T \boldsymbol{\Omega} (\boldsymbol{Y} - \boldsymbol{AX})|^{\frac{n}{2}}}{(2\pi)^{\frac{np}{2}}} \exp[-\frac{1}{2} \operatorname{Tr}[(\boldsymbol{Y} - \boldsymbol{AX})^T \boldsymbol{\Omega} (\boldsymbol{Y} - \boldsymbol{AX})]], \quad (4)$$

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

Obviously, one can propose many possible normalizations of the Gaussian likelihood as a function of the pair $(\boldsymbol{A}, \boldsymbol{\Omega})$. Our particular "choice" in Definition 1 is motivated by finding a function that ensures a finite maximum of the likelihood while keeping the form of the Gaussian density. This normalization of the Gaussian likelihood avoids exact solutions and subsequent divergence issues. The pair $(\boldsymbol{A}, \boldsymbol{\Omega})$ can then be computed to maximize the normalized-likelihood, L_N, i.e.,

$$(\boldsymbol{A}^*, \boldsymbol{\Omega}^*) = \arg\max_{\boldsymbol{A}, \boldsymbol{\Omega}} L_N(\boldsymbol{A}, \boldsymbol{\Omega}), \quad (5)$$

Proposition 1. *The solution to (5) is given by*

$$(\boldsymbol{Y} - \boldsymbol{A}^* \boldsymbol{X})^T \boldsymbol{\Omega}^* (\boldsymbol{Y} - \boldsymbol{A}^* \boldsymbol{X}) = n\boldsymbol{I}, \quad (6)$$

where $\boldsymbol{I}$ denotes the $n \times n$ Identity matrix.

Proof of Proposition 1: Let $\boldsymbol{Z} = (\boldsymbol{Y} - \boldsymbol{AX})^T \boldsymbol{\Omega} (\boldsymbol{Y} - \boldsymbol{AX})$. Then, the normalized-likelihood can be written as the following function of the variable $\boldsymbol{Z}$,

$$L_N(\boldsymbol{Z}) = \frac{|\boldsymbol{Z}|^{\frac{n}{2}}}{(2\pi)^{\frac{nq}{2}}} \exp -\frac{1}{2} \operatorname{Tr}[\boldsymbol{Z}]. \quad (7)$$

To find the stationary point $\boldsymbol{Z}^*$, we set $\dfrac{\partial L_N(\boldsymbol{Z})}{\partial \boldsymbol{Z}} = \boldsymbol{0}$.

$$\begin{aligned} \frac{\partial L_N(\boldsymbol{Z})}{\partial \boldsymbol{Z}} &= \frac{n}{2} |\boldsymbol{Z}|^{\frac{n}{2}-1} |\boldsymbol{Z}| \boldsymbol{Z}^{-1} \exp -\frac{1}{2} \operatorname{Tr}[\boldsymbol{Z}] \\ &- \frac{1}{2} |\boldsymbol{Z}|^{\frac{n}{2}} \exp -\frac{1}{2} \operatorname{Tr}[\boldsymbol{Z}] \\ &= \frac{1}{2} |\boldsymbol{Z}|^{\frac{n}{2}} [n\boldsymbol{Z}^{-1} - \boldsymbol{I}] \exp -\frac{1}{2} \operatorname{Tr}[\boldsymbol{Z}] \\ &= \boldsymbol{0} \\ &\Rightarrow \quad \boldsymbol{Z}^* = n\boldsymbol{I}. \end{aligned} \quad (8)$$

Moreover, it can be easily derived that the Hessian at the stationary point $\boldsymbol{Z}^*$ is given by

$$\frac{\partial^2 L_N(\boldsymbol{Z})}{\partial \boldsymbol{Z}^2} \mid_{\boldsymbol{Z}=\boldsymbol{Z}^*} = -\frac{1}{2n} n^{\frac{n^2}{2}} e^{-\frac{n^2}{2}} < 0 \tag{9}$$

■

There are many pairs $(\boldsymbol{A}^*, \boldsymbol{\Omega}^*)$, which satisfy equality (6) and hence maximize the normalized-likelihood. The non-uniqueness of the solution is not surprising given that the problem is under-determined. Among all possible solutions of (6), we propose to find those that minimize the regularized error $\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{\Omega}\|_F^2$, where λ is a tuning parameter and $\| \cdot \|_F$ denotes the Frobenius norm. Observe that it is meaningful to consider the error as the objective function here, because the set of pairs $(\boldsymbol{A}, \boldsymbol{\Omega})$ satisfying (6) are not exact solutions, i.e., they do not satisfy the equality $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, and hence the minimum error is not trivially zero. Thus, an advantage of the normalized-likelihood is that it avoids considering exact solutions. In addition, we consider constraints on the regression matrix $\boldsymbol{A}$, which reflect prior knowledge about the nature of the regression model. For instance, $\boldsymbol{A}$ may be constrained to be sparse. Many applications assume a sparse regression matrix, e.g., robust face recognition, where the target can be represented as a sparse linear combination of the dataset [9] and structural equation models (SEM) to infer gene or phenotype networks [10]. For now, let us consider a general constraint set, $\boldsymbol{A} \in \mathcal{A} \subset \mathbb{R}^{q \times p}$. The constrained optimization problem, thus, becomes

$$\begin{cases} \min\limits_{(\boldsymbol{A}, \boldsymbol{\Omega})} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{\Omega}\|_F^2 \\ \text{s.t.} \quad (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})^T \boldsymbol{\Omega} (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}) = n\boldsymbol{I}, \\ \qquad \boldsymbol{A} \in \mathcal{A}. \end{cases} \tag{10}$$

Problem (10) is formulated in terms of the two coupled variables $\boldsymbol{A}$ and $\boldsymbol{\Omega}$, which satisfy (6) to maximize the normalized-likelihood function. The following lemma derives an analytical expression of $\boldsymbol{\Omega}$ as a function of $\boldsymbol{A}$, and hence reduces the problem to depend on only one variable $\boldsymbol{A}$. Before stating the lemma's result, we need the following definition of the polar decomposition of matrices.

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

$$\boldsymbol{B} = \boldsymbol{U} |\boldsymbol{B}|, \tag{11}$$

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

Lemma 1. *Given $\boldsymbol{A}$, there exist many $\boldsymbol{\Omega}$ satisfying equality (6). The minimum Frobenius norm $\boldsymbol{\Omega}$, for a fixed $\boldsymbol{A}$, is given by*

$$\boldsymbol{\Omega}_A = n\, \boldsymbol{U} \left[(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})^T (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}) \right]^{-1} \boldsymbol{U}^T, \tag{12}$$

where $\boldsymbol{U}$ is the isometry of the matrix $(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})$.

Proof of Lemma 1: Let $\boldsymbol{B} = (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})$. Consider the polar decomposition of $\boldsymbol{B}$ given by

$$\boldsymbol{B} = \boldsymbol{U} |\boldsymbol{B}|, \quad \text{and} \quad |\boldsymbol{B}| = (\boldsymbol{B}^T \boldsymbol{B})^{1/2}. \tag{13}$$

Then, the equality in (6) becomes

$$\begin{aligned} & (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})^T \boldsymbol{\Omega} (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}) = n\boldsymbol{I} \\ \iff \; & \boldsymbol{B}^T \boldsymbol{\Omega} \boldsymbol{B} = n\boldsymbol{I} \\ \iff \; & |\boldsymbol{B}| \boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U} |\boldsymbol{B}| = n\boldsymbol{I} \\ \iff \; & \boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U} = n |\boldsymbol{B}|^{-2} \end{aligned} \tag{14}$$

Since $\boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{U}^T$ restricted to the range of $\boldsymbol{B}$ is invertible, i.e., $\boldsymbol{U}^T \restriction_{\text{Range}(\boldsymbol{B})}$ is invertible. Let us write

$$\mathbb{C}^q = \text{Range}(\boldsymbol{B}) \oplus \text{Ker}(\boldsymbol{B}^T), \tag{15}$$

where $\oplus$ denotes the direct sum of the two subspaces Range$(\boldsymbol{B})$ and Ker$(\boldsymbol{B}^T)$. Let $\boldsymbol{P}_B$ be the orthogonal projection onto Range$(\boldsymbol{B})$. Then, we can decompose $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \boldsymbol{P}_B \Omega \boldsymbol{P}_B \oplus \boldsymbol{P}_B \Omega \boldsymbol{P}_{B^\perp} \oplus \boldsymbol{P}_{B^\perp} \Omega \boldsymbol{P}_B \oplus \boldsymbol{P}_{B^\perp} \Omega \boldsymbol{P}_{B^\perp}, \tag{16}$$

where $\boldsymbol{P}_B^\perp$ is the orthogonal projection onto the orthogonal space of Range$(\boldsymbol{B})$, i.e., Ker$(\boldsymbol{B}^T)$. Recall that, by definition of the isometry $\boldsymbol{U}$, it satisfies the following properties:

$$\boldsymbol{P}_{B^\perp} \boldsymbol{U} = \boldsymbol{U}^T \boldsymbol{P}_{B^\perp} = \boldsymbol{0}. \tag{17}$$

Thus, from the decomposition of the matrix Ω in Eq. (16), we obtain

$$\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U} \quad = \quad \boldsymbol{U}^T \boldsymbol{P}_B \; \boldsymbol{\Omega} \; \boldsymbol{P}_B \boldsymbol{U}. \tag{18}$$

From Eq. (14) and since $\boldsymbol{U}^T \restriction_{\text{Range}(\boldsymbol{B})}$ is invertible, we have

$$\begin{aligned} & \boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U} = \boldsymbol{U}^T \boldsymbol{P}_B \; \boldsymbol{\Omega} \; \boldsymbol{P}_B \boldsymbol{U} = n |\boldsymbol{B}|^{-2} \\ \iff \; & \\ & \boldsymbol{P}_B \; \boldsymbol{\Omega} \; \boldsymbol{P}_B = n \; \boldsymbol{U} |\boldsymbol{B}|^{-2} \boldsymbol{U}^T. \end{aligned} \tag{19}$$

From the matrix decomposition in (16), for a fixed $\boldsymbol{A}$, $\boldsymbol{P}_B \; \boldsymbol{\Omega} \; \boldsymbol{P}_B$ is fixed. Thus, the minimum Frobenius norm matrix $\boldsymbol{\Omega}$ results by setting the three other terms in the matrix decomposition to zero, i.e., the minimum Frobenius norm matrix is of the form

$$\boldsymbol{\Omega} = \boldsymbol{P}_B \; \boldsymbol{\Omega} \; \boldsymbol{P}_B. \tag{20}$$

The result of Lemma 1 then follows from Eqs. (19) and (20).

■

Using Lemma 1, the following proposition states the equivalent form of problem (10), where the optimization problem does not depend on the variable $\boldsymbol{\Omega}$.

Proposition 2. *The optimization problem in (10) is equivalent to*

$$\begin{cases} \min\limits_{\boldsymbol{S}} \text{Tr}(\boldsymbol{S}^2) + \lambda \; n^2 \; \text{Tr}(\boldsymbol{S}^{-4}) \\ \textit{s.t.} \quad \boldsymbol{S} = |\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}|, \;\; \boldsymbol{A} \in \mathcal{A} \end{cases} \tag{21}$$

Proof of Proposition 2: Replacing $\boldsymbol{\Omega}_{\boldsymbol{A}}$ in the objective function of the optimization problem (10) by its expression obtained in Lemma 1, and letting $\boldsymbol{B} = \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}$, we obtain

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{\Omega}\|_F^2 \;&=\; \|\boldsymbol{B}\|_F^2 + \lambda \|n^2 \boldsymbol{U} (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{U}^T\|_F^2 \\ &=\; \text{Tr}(\boldsymbol{B}^T \boldsymbol{B}) + \lambda n^2 \\ &\quad\; \text{Tr}(\boldsymbol{U} (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{U}^T \boldsymbol{U} (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{U}^T) \\ &=\; \text{Tr}(\boldsymbol{B}^T \boldsymbol{B}) + \lambda n^2 \, \text{Tr}((\boldsymbol{B}^T \boldsymbol{B})^{-2}) \\ &=\; \text{Tr}(\boldsymbol{S}^2) + \lambda n^2 \, \text{Tr}(\boldsymbol{S}^{-4}), \end{aligned} \tag{22}$$

where $\boldsymbol{S}^2 = \boldsymbol{B}^T\boldsymbol{B} = (\boldsymbol{Y} - \boldsymbol{AX})^T(\boldsymbol{Y} - \boldsymbol{AX})$. ∎

Though the objective function in (21) is convex (as a function of the variable $\boldsymbol{S}$), the equality in the constraint (assuming that $\mathcal{A}$ is convex) is not affine and thus the optimization problem (21) is not convex [11]. We will, therefore, relax the minimization of (21) to a minimization over a convex set that is included in the original set. In what follows, we show that if the matrix regression $\boldsymbol{A}$ is sparse with a bounded norm, i.e., $\mathcal{A} = \{\boldsymbol{A} : \|A\|_1 \leq \epsilon\}$, then (21) can be approximated by a convex optimization problem. Moreover, this approximation formulates a much simpler optimization problem than the initial setting in (21) because it depends only on $\boldsymbol{S}$ and is independent of $\boldsymbol{A}$.

Proposition 3. *If $\mathcal{A} = \{\boldsymbol{A} : \|A\|_1 \leq \epsilon\}$, then the optimization problem in (21) can be approximated by the following convex optimization problem*

$$\begin{cases} \min\limits_{\boldsymbol{S}} \operatorname{Tr}(\boldsymbol{S}^2) + \lambda\, n^2 \operatorname{Tr}(\boldsymbol{S}^{-4}) \\ \text{s.t.} \quad \boldsymbol{S} \in \Lambda = \{\boldsymbol{S} \in \mathbb{S}_{n,n} \,:\, \|\boldsymbol{S} - |\boldsymbol{Y}|\|_F \leq \epsilon c^*\} \end{cases} \tag{23}$$

where $\mathbb{S}_{n,n}$ is the set of $n \times n$ symmetric positive semi-definite matrices and c^* is a small term which depends on X, Y but independent of ϵ.

Proof of Proposition 3: Let

$$\mathcal{S}_1 = \{\boldsymbol{S} : \boldsymbol{S} = |\boldsymbol{Y} - \boldsymbol{AX}|, \|\boldsymbol{A}\|_1 \leq \epsilon\}. \tag{24}$$

and let

$$\mathcal{S}_2 = \{\boldsymbol{S} \in \mathbb{S}_{n,n} : \|\boldsymbol{S} - |\boldsymbol{Y}|\|_F \leq \epsilon c^*\}. \tag{25}$$

We will show that $\mathcal{S}_2 \subseteq \mathcal{S}_1$. An illustration of these two sets is provided in Fig. 1. To this aim, we consider $\boldsymbol{S} \in \mathcal{S}_2$ and show that $\boldsymbol{S} \in \mathcal{S}_1$. Specifically, given $\boldsymbol{S} \in \mathcal{S}_2$ we find $\boldsymbol{A}$, such that $\boldsymbol{S} = |\boldsymbol{Y} - \boldsymbol{AX}|$ and $\|\boldsymbol{A}\|_1 \leq \epsilon$.

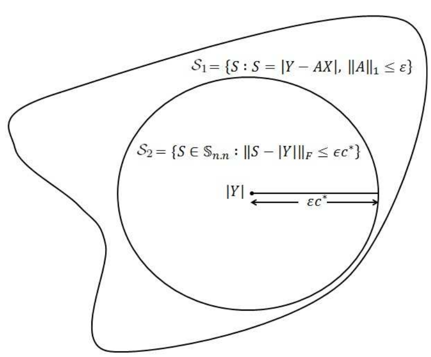


Fig. 1. Approximation of the optimization problem in Proposition 2 by the convex optimization problem in Proposition 3. Illustration of the sets $\mathcal{S}_1$ and $\mathcal{S}_2$ in the proof of Proposition 3.

Given $\boldsymbol{S} \in \mathbb{S}_{n,n}$, we want to find $\boldsymbol{A}$ such that $\boldsymbol{S} = |\boldsymbol{Y} - \boldsymbol{AX}|$, i.e., for some isometry $\boldsymbol{U}$ we have $\boldsymbol{US} = \boldsymbol{Y} - \boldsymbol{AX}$. For every isometry $\boldsymbol{U}$, one can find corresponding matrix $\boldsymbol{A}$ satisfying the previous identity. We will construct a specific matrix $\boldsymbol{A}$. Namely, we fix $\boldsymbol{U} = \boldsymbol{V}$, where $\boldsymbol{V}$ is the isometry from the polar decomposition $\boldsymbol{Y} = \boldsymbol{V}|\boldsymbol{Y}|$. Then, we need to find $\boldsymbol{A}$ such that

$$\boldsymbol{AX} = \boldsymbol{V}(|\boldsymbol{Y}| - \boldsymbol{S}). \tag{26}$$

$\boldsymbol{X}$ is full-rank; hence invertible from the right. Let us define

$$\tilde{\boldsymbol{X}} = \begin{cases} \boldsymbol{X}^{-1}\big|_{Range(X)}, \\ \boldsymbol{0}\big|_{[Range(X)]^\perp} \end{cases} \tag{27}$$

From the Definition of $\tilde{\boldsymbol{X}}$, we have $\boldsymbol{AX}\tilde{\boldsymbol{X}}\Big|_{Range(X)} = \boldsymbol{A}\Big|_{Range(X)}$ and $\boldsymbol{AX}\tilde{\boldsymbol{X}}\Big|_{Range(X)^\perp} = \boldsymbol{0}$. Therefore, multiplying Eq. (26) to the right by $\tilde{\boldsymbol{X}}$, we see that $\boldsymbol{A}$ defined by

$$\boldsymbol{A} = \begin{cases} \boldsymbol{V}(|\boldsymbol{Y}| - \boldsymbol{S})\tilde{\boldsymbol{X}}\Big|_{Range(X)}, \\ \boldsymbol{0}\Big|_{[Range(X)]^\perp} \end{cases} \tag{28}$$

solves Eq. (26). Now we estimate $\|A\|_1$,

$$\begin{aligned} \|\boldsymbol{A}\|_1 &\leq \|\boldsymbol{V}\|(\||\boldsymbol{Y}| - \boldsymbol{S}\|_1)\|\tilde{\boldsymbol{X}}\| \\ &\leq n\|\boldsymbol{V}\|(\||\boldsymbol{Y}| - \boldsymbol{S}\|_F)\|\tilde{\boldsymbol{X}}\| && (29) \\ &= C'\||\boldsymbol{Y}| - \boldsymbol{S}\|_F && (30) \\ &\leq C'\epsilon c^* && (31) \end{aligned}$$

where (29) follows from the equivalence of norms and Cauchy-Schwartz. In (30), $C' = n\|\boldsymbol{V}\|\|\tilde{\boldsymbol{X}}\|$, which is a constant. The inequality in (31) follows from the fact that $\boldsymbol{S} \in \mathcal{S}_2$ and $\|\boldsymbol{S} - |\boldsymbol{Y}|\|_F \leq \epsilon c^*$. In (31), by choosing $c^* \leq \frac{1}{C'} = 1/(n\|\boldsymbol{V}\|\|\tilde{\boldsymbol{X}}\|)$, we obtain $\boldsymbol{A} \leq \epsilon$. This ends the proof that $\boldsymbol{S} \in \mathcal{S}_1$. ∎

The optimization problem (23) is convex, hence it admits a unique global solution $\boldsymbol{S}^*$. Given $\boldsymbol{S}^*$, the optimal regression matrix, $\hat{\boldsymbol{A}}$, is found by solving $\boldsymbol{S}^* = |\boldsymbol{Y} - \hat{\boldsymbol{A}}\boldsymbol{X}|$. There are many possible such solutions $\hat{\boldsymbol{A}}$. We propose to find the sparsest matrix, in the sense of minimization of the l_1-norm.

$$\begin{cases} \min\limits_{\boldsymbol{A},\boldsymbol{U}} \|\boldsymbol{A}\|_1 \\ \text{s.t.} \quad \boldsymbol{AX} = \boldsymbol{Y} - \boldsymbol{US}^*, \end{cases} \tag{32}$$

where $\boldsymbol{U}$ is an isometry matrix. For every isometry $\boldsymbol{U}_0$, we can find the minimum l_1-norm $\boldsymbol{A}(\boldsymbol{U}_0)$. The optimal matrix $\boldsymbol{A}$ is, thus, found by minimizing over $\boldsymbol{U}$ and $\boldsymbol{A}$. Let $\boldsymbol{V}$ be the isometry of the matrix $\boldsymbol{Y}$. Assuming that $\boldsymbol{A}$ is sparse, we can chose $\boldsymbol{U}$ to be the isometry of $\boldsymbol{Y}$. By replacing $\boldsymbol{U}$ by $\boldsymbol{V}$ in (32), we may increase the minimum but we reduce the problem to a convex setting in the unique variable $\boldsymbol{A}$. Finally, the estimated regression matrix is the unique global solution of the following convex optimization problem,

$$\begin{cases} \min\limits_{\boldsymbol{A}} \|\boldsymbol{A}\|_1 \\ \text{s.t.} \quad \boldsymbol{AX} = \boldsymbol{Y} - \boldsymbol{VS}^*, \end{cases} \tag{33}$$

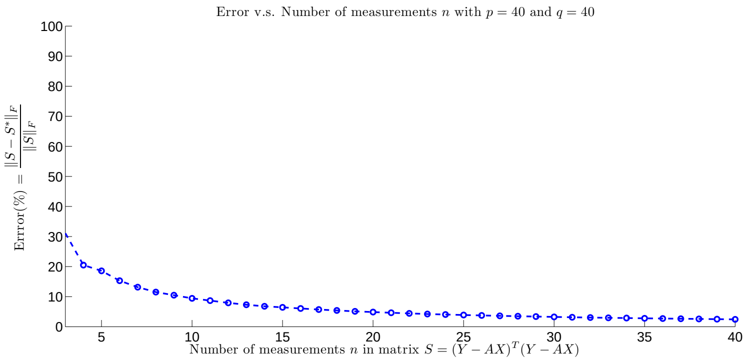


Fig. 2. Approximation error $||\boldsymbol{S} - \boldsymbol{S}*||_F/||\boldsymbol{S}||_F$ versus $n = 1, \cdots, p$ for $p = 40$.

SMURC algorithm

The SMURC algorithm is summarized below.

Input: The matrices $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{q\times n}$ according to the multivariate regression model in Eq. (2) with $q > n$.

Step 1 Solve the convex optimization problem in (23). The solution of this problem is a p.s.d. matrix $\boldsymbol{S}^* \in \mathbb{R}^{n\times n}$

Step 2 Given $\boldsymbol{S}^*$, the optimal regression matrix is obtained as the solution to the convex optimization problem in (33).

Steps 1 and 2 can be implemented efficiently using the Matlab Software for Disciplined Convex Programming, *cvx* [12], [13].

The following corollary provides an upper bound on the l_1-norm of the optimal connectivity matrix

Corollary 1. *The norm of the optimal connectivity matrix, given by the solution of the convex optimization problem in (33), is bounded above by*

$$\|\boldsymbol{A}^*\|_1 \leq \|\boldsymbol{V}(|\boldsymbol{Y}| - \boldsymbol{S}^*)\tilde{\boldsymbol{X}}\|_1 \leq \epsilon, \tag{34}$$

where $\boldsymbol{V}$ is the isometry in the polar decomposition of $\boldsymbol{Y}$, $\boldsymbol{S}^*$ is the global solution of the convex optimization problem in (23) and $\tilde{\boldsymbol{X}}$, defined in (27), is the right inverse of the matrix $\boldsymbol{X}$.

Proof: The proof follows from the proof of Proposition 3, and specifically from Eq. (29). ■

The SMURC algorithm involved an approximation of the original optimization problem (10) by the convex optimization problem in (23). It is thus important to assess the effect of this convex approximation on the final solution. An analytical derivation to bound this approximation is difficult and cumbersome. We, therefore, provide a numerical assessment of this approximation by computing the average error between the exact solution of (21) and the approximate solution of (23), $||\boldsymbol{S} - \boldsymbol{S}*||_F/||\boldsymbol{S}||_F$. In synthetic data, the exact solution $\boldsymbol{S}$ is known. The error graph, displayed in Fig. 2 shows that this approximation error decreases to a very small value when the number of measurements n approaches the number of unknowns p.

III. Application: Genetic Regulatory Networks

An application of interest, which suffers from the high-dimension, small sample-size problem is the reconstruction, also called *reverse engineering*, of genetic regulatory networks (GRNs), where only few samples, denoting time points or tissue samples, are available. Inference of genetic regulatory networks is important for understanding the dynamics of genetic interactions and harnessing this understanding into an educated intervention of the cell. The behavior of the regulatory network can be modeled by a system of linear differential equations near a steady-state [14]–[18]:

$$\dot{x}_i(t_k) = \sum_{j=1}^{N} a_{ij}x_j(t_k) + b_i u(t_k) + \epsilon_i(t_k), \tag{35}$$

where $i = 1, \cdots, p, k = 1, \cdots, n$, p is the number of genes, n 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. $\epsilon_i(t_k)$ models the measurement and model error at time step k. The goal is to infer the gene interactions $\{a_{ij}\}_{i,j=1}^{p}$, given a certain number of measurements n. Introducing the new variable $y_i(t) = \frac{dx_i}{dt} - b_i u(t)$, we can write the ODE model in vector form for the p genes as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{36}$$

where $\boldsymbol{y} = [y_1, y_2, \cdots, y_p]^T, \boldsymbol{x} = [x_1, x_2, \cdots, x_p]^T, \boldsymbol{\epsilon} = [\epsilon_1, \cdots, \epsilon_p]^T$ and $\boldsymbol{A} = \{a_{ij}\}_{i,j=1}^{p}$. Performing n different experiments , we obtain n measurements and can write the results as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}, \tag{37}$$

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

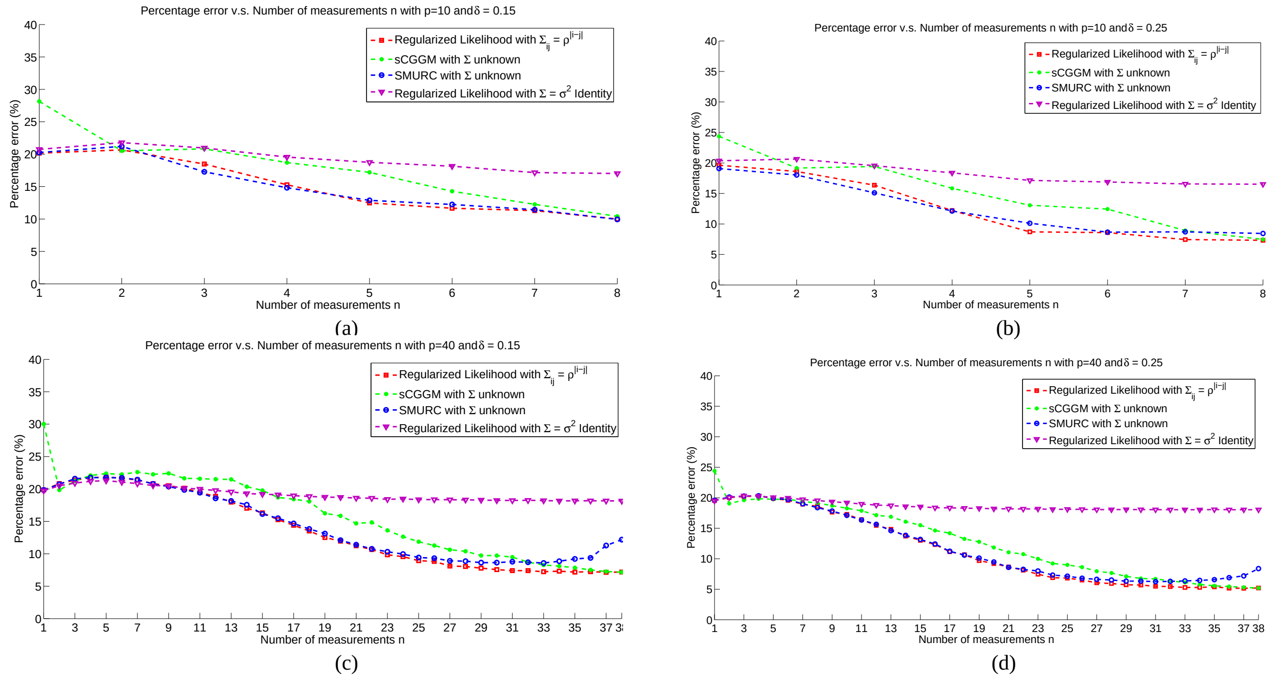


Fig. 3. Performance comparison of SMURC with sCGGM and the l_1-regularized maximum likelihood (RMLE) with known covariance for different network sizes with %80 sparsity. Blue: SMURC with unknown covariance; Green: sCGGM with unknown covariance; Red: RMLE with $\Sigma = \Sigma_{true} = \rho^{|i-j|}$; Purple: RMLE with $\Sigma = \sigma^2 I$, where σ^2 is estimated from the data. (a) $(p,\delta) = (10, 0.15)$; (b) $(p,\delta) = (10, 0.25)$; (c) $(p,\delta) = (40, 0.15)$; (d) $(p,\delta) = (40, 0.25)$.

A. Simulation results

Before considering a real dataset, we generate synthetic data and compare the proposed SMURC algorithm with the l_1-regularized maximum likelihood estimator in [14], where an l_1-norm penalty is imposed on the connectivity matrix $\boldsymbol{A}$. The regularized MLE finds the optimal connectivity matrix, which minimizes the following convex function

$$f(\boldsymbol{A}) = \mathrm{Tr}\left[\frac{1}{n}(\boldsymbol{Y} - \boldsymbol{AX})(\boldsymbol{Y} - \boldsymbol{AX})^T\boldsymbol{\Sigma}^{-1}\right] + \ln|\boldsymbol{\Sigma}| + \alpha\sum_{i=1}^{p}\sum_{j=1}^{p}|a_{i,j}|, \tag{38}$$

where $\boldsymbol{\Sigma}$, the covariance matrix of the data, is assumed to be known and α is a tuning parameter that controls the sparsity level of the matrix $\boldsymbol{A}$.

We generate synthetic gene networks with varying size p, varying number of measurements $n < p$, and covariance structure $\boldsymbol{\Sigma}$. Gene regulatory networks are known to be sparse: every gene interacts only with few other genes. Thus, the connectivity matrix $\boldsymbol{A}$ is sparse. In the presented simulations, we assume 80% sparsity level, i.e., $\|\boldsymbol{A}\|_0 = 0.2p^2$, where $\|\cdot\|_0$ denotes the number of non-zero elements. The performance of the algorithm is similar for other sparsity levels as long as the system is under-determined. The entries of the matrix $\boldsymbol{A}$ are drawn from a standard normal distribution with zero-mean and unit variance, i.e., $a_{i,j} \sim \mathcal{N}(0,1)$. We use the same covariance matrix suggested in [14], [19], $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.7$. The performance of the algorithm is assessed using the following measure suggested in [19]:

$$E = \sum_{i=1}^{p}\sum_{i=j}^{p} e_{i,j} \quad \text{with}$$

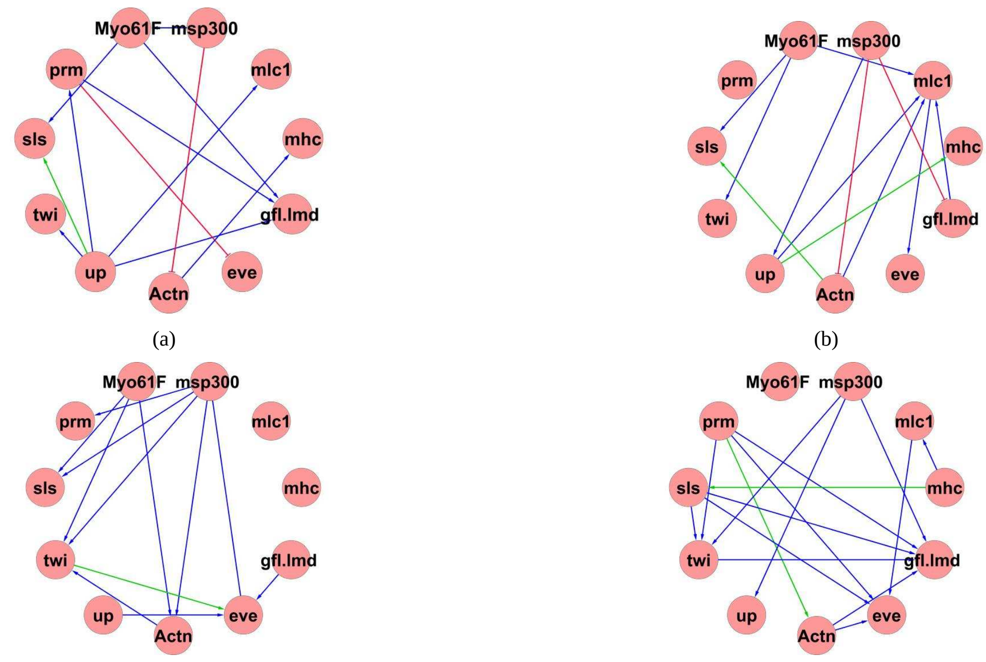
$$e_{i,j} = \begin{cases} 1, & \text{if} \quad |a_{ij} - \hat{a}_{ij}| > \delta|a_{ij}| \\ 0, & \text{otherwise}, \end{cases} \tag{39}$$

where a_{ij} is the $(i,j)^{th}$ element of the true genetic interaction matrix and $\hat{a}_{ij}$ is the estimate of a_{ij}. δ is a threshold parameter. The percentage error is computed as E/p^2.

Figure 3 shows the percentage error versus the number of measurements n for $p = 10$ and $p = 40$-gene networks, which are 80% sparse. We considered a threshold of error corresponding to $\delta = 0.15$ and $\delta = 0.25$. Observe that, though the system is sparse, it is still under-determined, i.e., the number of "effective" unknowns is larger than the number of independent observations. We compare the proposed SMURC algorithm (which assumes an unknown covariance matrix) with the sCGGM algorithm [8] and the regularized MLE with the true covariance matrix [14] and with a diagonal covariance matrix $\Sigma = \sigma^2 I$, where σ^2 is estimated from the data. It was shown in [8] that sCGGM outperforms Rothman *et al.* MRCE and approximate MRCE. We used the optimized code for sCGGM available at http://www.cs.cmu.edu/~sssykim/softwares/softwares.html. We assess the algorithms with a covariance $\boldsymbol{\Sigma}_{true} = \rho^{|i-j|}$ with $\rho = 0.7$. Fifty Monte Carlo simulations were performed for each experiment.

B. Drosophila Melanogaster gene expression data

To assess our algorithm on real data, we tested it on the Drosophila melanogaster gene expression levels [21]. The data contains 4028 genes in wild-type flies examined during 66



(a)

(b)

(c)

(d)

Fig. 5. Estimated gene regulatory networks of the Drosophila during four developmental phases using the SMURC algorithm. Blue and red edges denote, respectively, positive and negative interactions. The green edges are interactions reported in Flybase. (a) Embryonic; (b) Larval; (c) Pupal; (d) Adulthood.

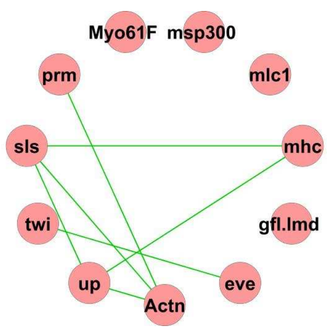


Fig. 4. Flybase: The known undirected gene interactions in the Drosophila's 11-gene wing muscle network [20].

sequential time periods beginning at fertilization and spanning embryonic, larval, pupal and the first 30 days of adulthood. Since early embryos change rapidly, overlapping 1-hour periods were sampled; adults were sampled at multiday intervals. The time points span the embryonic (samples 1-30; time E01h till E2324h), larval (samples 31-40; time L24h till L105h), pupal (samples 41-58; M0h till M96h) and adulthood (samples 59-66; A024h till A30d) periods of the organism. A list of known undirected gene interactions is hosted in Flybase [20].

A set of 11 genes that regulate the wing muscle development has been considered in [22]–[25]. The 11-gene network, with the interactions reported in Flybase, is depicted in Fig. 4. We reconstructed the genetic network between these 11 genes during the four developmental phases using the SMURC algorithm. In the embryonic and pupal phases, 9 time points, undersampled from the original time points (30 for embryonic and 18 for pupal), were used to reconstruct the 11-gene network during these two developmental periods. In the larval and adulthood phases, the entire 9 larval and 7 adulthood time points were used to reconstruct the network during the larval and adulthood development phases, respectively. In summary, the connectivity matrix of the 11-gene Drosophila development network was estimated using the SMURC algorithm with 9 time points in the embryonic phase, 9 time points in the larval phase, 9 time points in the pupal phase and 7 time points in the adulthood phase. Observe that in all four developmental phases, the system is underdetermined.

The reconstructed networks using the SMURC algorithm are shown in Fig. 5. The SMURC algorithm was able to detect six out of the seven Flybase interactions during different developmental phases of the organism: (*up*,*sls*) appears during the embryonic period; (*Actn*,*sls*) and (*up*,*mhc*) appear during the larval phase; (*twl*,*eve*) appears during the pupal phase; (*prm*,*Actn*) and (*mhc*,*sls*) appear during the adulthood stage of the development.

We compare the SMURC findings with the results in [22], [23], [24], [25]. Though these references are not directly related to the problem of under-determined regression systems with unknown covariance structure, their work aims at reverse-engineering the connectivity of genetic regulatory networks. In particular, they all consider the Drosophila's 11-gene wing muscle network. Zhao *et al.* [22] infer a single directed network using the minimum description length principle. They used all 66 time points to identify a single network that characterizes the entire Drosophila's life cycle. In [23], a time-varying undirected network is learnt using an exponential random graph model model. A dynamic Bayesian network was used in [24], and [25] proposed a non-parametric Bayesian regression approach for gene regulatory network inference.

Table I shows the detection of the known interactions in Flybase by the five approaches, E,L,P,A stand, respectively, for the embryonic, larval, pupal and adulthood phases. Though the proposed SMURC algorithm relies on fewer time points than the other approaches, it detected the most number of known interactions cited in Flybased and reported in FLIGHT website http://flight.icr.ac.uk/search/search_interactions.jsp. Additionally, the SMURC algorithm found two directed interactions (*msp 300* $\rightarrow$ *prm*) and (*msp300* $\rightarrow$ *up*) in common with the works in [22], [23], [24], and three directed interactions during the embryonic phase in common with [25] (the networks in the other phases were not reported in [25]) , (*up* $\rightarrow$ *twi*), (*up* $\rightarrow$ *mlc1*) and (*msp300* $\longrightarrow$ *Myo61F1*). It is striking that all detected interactions that are shared with previous work [22]–[25] have also the same direction.

IV. Conclusion and Discussion

In this paper, we showed that the Gaussian likelihood, as a function of the regression coefficients and the covariance matrix, diverges when the multivariate regression system is under-determined. We subsequently proposed a normalized likelihood function that guarantees convergence while still keeping the Gaussian form of the data. The maximum normalized likelihood, however, admits multiple solutions because the system is still under-determined. Using the polar decomposition of matrices, we provided an expression of the covariance matrix in terms of the regression coefficients. This provided an equivalent representation of the optimization problem in one variable only, namely the regression matrix. We then relaxed the optimization problem into a convex one by considering a convex set included in the original constraint set. The optimal sparse regression matrix is found as the global solution to a convex optimization problem.

We applied the proposed Small-sample MUltivariate Regression with Covariance estimation (SMURC) algorithm to infer the wing muscle genetic regulatory networks of the Drosophila melanogaster during the four phases of its development: embryonic, larval, pupal and adulthood. Genetic regulatory networks are known to be sparse and often the number of measurements is smaller than the number of genes, which makes the network inference problem unidentifiable. SMURC was able to detect six out of the seven interactions reported in Flybase. Other algorithms aimed at reverse-engineering dynamic gene regulatory networks were able to detect a maximum of three out of the seven interactions.

Acknowledgment

This project is supported by Award Number R01GM096191 from the National Institute Of General Medical Sciences (NIH/NIGMS) and Award Number ACI-1429467 from the National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the funding agencies.

References

[1] O. Hauk, “Keep it simple: a case for using classical minimum norm estimation in the analysis of EEG and MEG data,” *Neuroimage*, vol. 21, no. 4, pp. 1612–1621, April 2004.

[2] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang, “Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer,” *Annals of Applied Statistics*, vol. 4, no. 1, pp. 53–77, 2010.

[3] F. D. Martinoa, A. W. de Borsta, G. Valentea, R. Goebela, and E. Formisanoa, “Predicting EEG single trial responses with simultaneous fmri and relevance vector machine regression,” *NeuroImage*, vol. 56, no. 2, pp. 826–836, May 2011.

[4] G. Reinsel and R. Velu, *Multivariate Reduced-rank Regression: Theory and Applications*. New York: Springer, 1998.

[5] M. Yuan and Y. Lin, “Model selection and estimation in the gaussian graphical model,” *Biometrika*, vol. 94, no. 1, p. 1935, 2007.

[6] N. B. Mohammed Mohammed-Rasheed and H. Fathallah-Shaykh, “A combined constraint approach for inference of sparse large-scale biomolecular networks,” in *International Conference on Control, Engineering & Information Technology*, vol. 1, 2013, pp. 38–42.

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

[8] L. Zhang and S. Kim, “Learning gene networks under SNP perturbations using eQTL datasets,” *PLoS Computationaly Biology 10(4)*, vol. 10, February 2014.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transactions on pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210 – 227, February 2009.

[10] X. Cai, J. A. Bazerque, and G. B. Giannakis, “Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations.” *PLoS Computational Biology*, vol. 9, no. 5, May 2013.

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

[12] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.0 beta,” http://cvxr.com/cvx, 2013.

[13] ——, “Graph implementations for nonsmooth convex programs,” in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[14] G. Rasool, N. Bouaynaya, H. M. Fathallah-Shaykh, and D. Schonfeld, “Inference of genetic regulatory networks using regularized likelihood with covariance estimation,” in *IEEE Statistical Signal Processing Workshop*, August 2012.

[15] H. M. Fathallah-Shaykh, J. L. Bona, and S. Kadener, “Mathematical model of the Drosophila circadian clock: Loop regulation and transcriptional integration,” *Biophysical Journal*, vol. 97, no. 9, pp. 2399–2408, November 2009.

[16] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo, “How to infer gene networks from expression profiles,” *Molecular Systems Biology*, vol. 3, no. 78, February 2007.

[17] 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, April 2006.

[18] M. 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.

[19] 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.

[20] S. J. Marygold, P. C. Leyland, R. L. Seal, J. L. Goodman, J. Thurmond, V. B. Strelets, R. J. Wilson *et al.*, “Flybase: improvements to the bibliography,” *Nucleic acids research*, vol. 41, no. D1, pp. 751–757, 2013.

[21] M. N. Arbeitman, E. E. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White, “Gene expression during the life cycle of drosophila melanogaster,” *Science*, vol. 297, no. 5590, pp. 2270–2275, 2002.

[22] W. Zhao, E. Serpedin, and E. R. Dougherty, “Inferring gene regulatory networks from time series data using the minimum description length principle,” *Bioinformatics*, vol. 22, no. 17, pp. 2129–2135, 2006.

TABLE I
DETECTION OF THE KNOWN GENE INTERACTIONS IN FLYBASE

	(*prm*,*Actn*)	(*sls*,*mhc*)	(*mhc*,*up*)	(*sls*,*Actn*)	(*sls*,*up*)	(*twi*,*eve*)	(*up*,*Actn*)
SMURC	✓ (A)	✓ (A)	✓ (L)	✓ (L)	✓ (E)	✓ (P)	×
Minimum description length [22]	✓	✓	×	×	×	✓	×
Random graph model [23]	×	×	✓ (E,L,P,A)	✓ (P,A)	✓ (E,L,P,A)	×	×
Dynamic Bayesian network [24]	×	✓ (E,L,P,A)	×	×	×	×	×
Nonparametric Bayesian regression [25]	×	×	×	×	×	✓ (E)	×

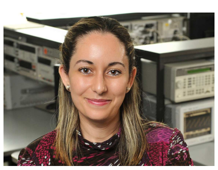
[23] F. Guo, S. Hanneke, W. Fu, and E. P. Xing, "Recovering temporally rewiring networks: A model-based approach," in *Proceedings of the international conference on Machine learning*, 2007, pp. 321–328.

[24] J. W. Robinson and A. J. Hartemink, "Non-stationary dynamic Bayesian networks," in *Advances in Neural Information Processing Systems*, 2008, pp. 1369–1376.

[25] H. Miyashita, T. Nakamura, Y. Ida, T. Matsumoto, and T. Kaburagi, "Nonparametric Bayes-based heterogeneous "drosophila melanogaster" gene regulatory network inference: T-process regression," in *Proceedings of the International Conference on Artificial Intelligence and Applications*, 2013, pp. 51–58.

Roman Shterenberg received the B.S. degree in Physics in 1998, the M.S. and Ph.D. degrees in Mathematics in 2000 and 2003 from St. Petersburg State University, Russia. In 2005-2007, he was a Van Vleck Assistant Professor at University of Wisconsin-Madison. In 2007, he joined The University of Alabama at Birmingham, where he is currently an Associate Professor in the Department of Mathematics. His research interests are in mathematical physics, spectral theory, inverse problems and mathematical biology.

Belhassen Bayar Belhassen Bayar received the B.S. degree in Electrical and Computer Engineering from the Ecole Nationale d'Ingénieurs de Tunis (ENIT), Tunisia, in 2011, and the MS degree in Electrical and Computer Engineering from Rowan University, New Jersey, in 2014. In Fall 2014, he joined Drexel University, Pennsylvania, where he is currently a PhD candidate with the Department of Electrical and Computer Engineering. Bayar won the Best Paper Award at the IEEE International Workshop on Genomic Signal Processing and Statistics in 2013. His main research interests are in signal processing, machine learning, and forensics.



Nidhal Bouaynaya received the B.S. degree in Electrical Engineering and Computer Science from the Ecole Nationale Supérieure de L'Electronique et de ses Applications (ENSEA), France, in 2002, the MS degree in Electrical and Computer Engineering from the Illinois Institute of Technology, Chicago, in 2002, the Diplôme d'Etudes Approfondies in Signal and Image processing from ENSEA, France, in 2003, the M.S. degree in Mathematics and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Chicago, in 2007. From 2007-2013, she was an Assistant then Associate Professor with the Department of Systems Engineering at the University of Arkansas at Little Rock. Since 2013, she joined Rowan University, where she is currently an Associate Professor with the Department of Electrical and Computer Engineering. Dr. Bouaynaya won the Best Student Paper Award in Visual Communication and Image Processing 2006, the Best Paper Award at the IEEE International Workshop on Genomic Signal Processing and Statistics 2013 and the runner-up Best Paper Award at the IEEE International Conference on Bioinformatics and Biomedicine 2015. She is currently serving as an Associate Editor for EURASIP Journal on Bioinformatics and Systems Biology. Her current research interests are in biomedical signal processing, medical imaging, mathematical biology and dynamical systems.