Closed-Form Solution and Sparsity Path for Inverse Covariance Estimation Problem

Salar Fattahi and Somayeh Sojoudi

***Abstract*— In this paper, we study the problem of determining a sparsity graph that describes the conditional dependence between different elements of a multivariate random distribution, using a limited number of samples. Graphical Lasso is one of the popular methods for addressing this problem, which imposes a soft penalty in the form of an l_1 regularization term in its objective function. The first goal of this work is to study the behavior of the optimal solution of Graphical Lasso as a function of its regularization coefficient. We show that if the number of samples is not too small compared to the number of parameters, the sparsity pattern of the optimal solution of Graphical Lasso changes gradually in terms of the regularization coefficient. More precisely, it is proved that each change in the sparsity pattern corresponds to the addition or removal of a single edge of the graph, under generic conditions. It is also shown that Graphical Lasso as a conic optimization problem has a closed-form solution if an acyclic graph is sought. This explicit formula also serves as an approximate solution for non-acyclic sparse graphs. The results are demonstrated on synthetic data and electrical systems.**

I. Introduction

There has been a pressing need in developing new and efficient computational methods to analyze and learn the characteristics of high-dimensional data with a structured or randomized nature. Real-world datasets are often overwhelmingly complex, and therefore it is important to obtain a simple description of the data that can be processed efficiently. In an effort to address this problem, there has been a great deal of interest in sparsity-promoting techniques for large-scale optimization problems [1], [2]. These techniques have become essential to the tractability of big-data analyses in many applications, including data mining [3], pattern recognition [4], [5], human brain functional connectivity [6], distributed controller design [7]–[9], multi-agent systems [10], and compressive sensing [11], [12]. Similar approaches have been used to arrive at a parsimonious estimation of high-dimensional data. Most of the statistical learning techniques in data analytics are contingent upon the availability of a sufficient number of samples (compared to the number of parameters), which is difficult to satisfy for many applications [13], [14]. To remedy the aforementioned issues, a special attention has been paid to the augmentation of various problems with sparsity-inducing penalty functions to obtain sparse and easy-to-analyze solutions. A Lasso-type penalty function is often used in the literature to estimate the inverse of the covariance matrix, which gives rise to the notion of Graphical Lasso [15], [16].

Recently, it has been shown in different applications, such as brain connectivity networks and electrical circuits, that the thresholding technique and Graphical Lasso would result in the same sparsity structure [17]. This result is significant as it leads to a major reduction in the computational complexity of Graphical Lasso. More precisely, it has been shown in [17] that the computationally-expensive GL problem and the simple thresholding approach result in the same support graph as long as the solution of the GL problem is close to its first-order Taylor approximation or, equivalently, the regularization coefficient is large enough. The conditions proposed in [17] depend on the optimal solution of the GL problem and, thus, they cannot be used in practice beyond a theoretical analysis. The paper [18] addresses this issue by deriving verifiable conditions under which the sparsity pattern of the GL solution can be found using the thresholding technique, but it cannot provide any information on the values of the entries of the GL solution. Another line of work has been devoted to studying the connectivity structure of the optimal solution of the GL problem. In particular, [19] and [20] have shown that the connected components induced by thresholding the sample covariance matrix and those induced by the optimal solution of the GL problem have the same vertex partitioning. Although this result does not require any particular condition, it cannot provide any information about the edge structure of the support graph. Therefore, one may need to solve GL for each connected component using an iterative algorithm, which may take up to $\mathcal{O}(d^3)$ operations per iteration for a problem with d random variables [15], [16], [19].

A. Problem Formulation

Consider a random vector $\mathcal{X} = [x_1, x_2, ..., x_d]$ with an underlying multivariate Gaussian distribution. Let Σ_* denote the correlation matrix of this random vector. Without loss of generality, we assume that $\mathcal{X}$ has a zero mean. The goal is to estimate the sparsity pattern and/or the entries of Σ_*^{-1} based on n independent samples $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, ..., \mathbf{x}_{(n)}$ of $\mathcal{X}$. The sparsity pattern of Σ_*^{-1} determines which random variables in $\mathcal{X}$ are conditionally independent. In particular, if the $(i,j)^{\text{th}}$ entry of Σ_*^{-1} is zero, it means that x_i and x_j are independent given the remaining entries of $\mathcal{X}$ (the value of this entry is proportional to the partial correlation between x_i and x_j). In this paper, we assume that Σ_*^{-1} is sparse and that Σ_* is non-singular. The problem of studying the conditional

Email: fattahi@berkeley.edu and sojoudi@berkeley.edu.

Salar Fattahi is with the Department of Industrial Engineering and Operations Research, University of California, Berkeley. Somayeh Sojoudi is with the Departments of Electrical Engineering and Computer Sciences and Mechanical Engineering, University of California, Berkeley. This work was supported by the ONR grant N00014-17-1-2933, DARPA grant D16AP00002, and AFOSR grant FA9550-17-1-0163.

independence of different entries of $\mathcal{X}$ is challenged in practice by the fact that the true correlation matrix is rarely known *a priori*. Therefore, the sample correlation matrix should instead be used to estimate the true correlation matrix. Let Σ denote the sample correlation matrix. To estimate Σ_*^{-1}, consider the optimization problem

$$\underset{S\in\mathbb{S}_+^d}{\text{minimize}} \ -\log\det(S)+\text{trace}(\Sigma S) \tag{1}$$

The optimal solution of the above problem is equal to $S^{\text{opt}}=\Sigma^{-1}$. However, there are two issues with this solution. First, the number of samples available in many applications is modest compared to the dimension of Σ. This makes Σ ill-conditioned or even singular, which would lead to large or undefined entries for the optimal solution of (1). Second, although Σ_*^{-1} is assumed to be sparse, a small random difference between Σ_* and Σ would make S^{opt} highly dense. In order to address the aforementioned issues, consider the problem

$$\underset{S\in\mathbb{S}_+^d}{\text{minimize}} \ -\log\det(S)+\text{trace}(\Sigma S)+\lambda\|S\|_* \tag{2}$$

where $\lambda>0$ is a regularization coefficient. This problem is known as Graphical Lasso (GL). The term $\|S\|_*$ in the objective function is defined as the summation of the absolute values of the off-diagonal entries in S. This additional penalty acts as a surrogate for promoting sparsity in the off-diagonal elements of S, while ensuring that the problem is well-defined even with a singular input Σ. The main goal of this paper is twofold:

1. First, we examine the behavior of the sparsity pattern of the optimal solution of Graphical Lasso with respect to the regularization coefficient λ. In particular, we show that the sparsity structure of the optimal solution changes gradually with respect to λ if the number of samples used to construct the sample covariance matrix is not too small. This implies that, using GL, any level of sparsity is achievable in the estimated inverse correlation matrix by fine-tuning the regularization coefficient via a simple bisection method.
2. Although (2) is a convex optimization problem, its scalability is limited to small- and medium-sized problems. To partially address this problem, we develop an explicit formula for the solution of Graphical Lasso if an acyclic graph is sought (see the technical report [21] for the generalization of this result to non-acyclic graphs).

The above results shed light on both the sparsification property of Graphical Lasso and its computational complexity. Although conic optimization problems almost never benefit from an explicit formula for their solutions and need to be solved numerically, the formula found in the paper suggests that sparse Graphical Lasso and related graph-based conic optimization problems may fall into the category of problems with closed-form solutions (e.g., least squares problems).

Although the Graphical Lasso technique is commonly used for estimating the inverse of the correlation matrix for Gaussian distributions, a similar estimation method can be used for random vectors with other probability distributions. This is due to the fact that this method corresponds to the minimization of the l_1-regularized log-determinant Bregman divergence, which is a widely used metric for measuring the distance between the true and estimated parameters of a problem [22]. Therefore, the theoretical results developed in this paper are applicable to more general learning problems.

Notations: Lowercase, bold lowercase and uppercase letters are used for scalars, vectors and matrices, respectively (say $x,\mathbf{x},X$). $\mathbb{R}^n$, $\mathbb{S}^n$, and $\mathbb{S}_+^n$ are used to denote the sets of $n\times 1$ real vectors, $n\times n$ symmetric matrices, and $n\times n$ symmetric and positive-semidefinite matrices, respectively. Furthermore, $\mathbb{S}_+^{n,r}$ denotes the set of $n\times n$ positive semidefinite matrices of rank at most r. The symbols $\text{trace}(M)$ and $\log\det(M)$ refer to the trace and the logarithm of the determinant of the matrix M, respectively. The $(i,j)^{\text{th}}$ entry of the matrix M is denoted by M_{ij}. Moreover, I_n denotes the $n\times n$ identity matrix. The sign of a scalar x is shown by $\text{sign}(x)$. The notations $|x|$ and $\|M\|_1$ denote the absolute value of the scalar x and the induced 1-norm of the matrix M, respectively. The inequality $M\succeq 0$ means that M is positive semidefinite. The cardinality of a discrete set $\mathcal{D}$ is denoted as $|\mathcal{D}|_0$. Given a matrix $M\in\mathbb{S}^d$, define

$$\|M\|_*=\sum_{i=1}^d\sum_{j=1}^d|M_{ij}|-\sum_{i=1}^d|M_{ii}| \tag{3}$$

$$\|M\|_{\max}=\max_{i\neq j}|M_{ij}| \tag{4}$$

II. Main Results

The first objective is to examine the behavior of the sparsity pattern of the solution of Graphical Lasso as a function of λ.

Definition 1: Given a matrix $M\in\mathbb{S}^d$, the **support or sparsity graph** of M, denoted by $\text{supp}(M)$, is defined as a graph with the vertex set $\mathcal{V}=\{1,2,...,d\}$ and the edge set $\mathcal{E}$, where $(i,j)\in\mathcal{E}$ if and only if $M_{ij}\neq 0$ and $i\neq j$.

Definition 2: Let $S^{\text{opt}}(\lambda)$ denote the unique symmetric optimal solution of (2) for a given λ. Define the active set of $S^{\text{opt}}(\lambda)$ as the set of nonzero elements of the upper triangular part of $S^{\text{opt}}(\lambda)$, which is denoted by $I_{ac}(\lambda)$. It is said that λ^* is a breakpoint if $I_{ac}(\lambda)$ changes at $\lambda=\lambda^*$. Denote the set of all breakpoints as Λ_b.

For notational simplicity, we use S^{opt} instead of $S^{\text{opt}}(\lambda)$ whenever the equivalence is implied by the context. The first main result of this paper is stated below.

Theorem 1: The following statements hold:

1. If $n\geq d$, the cardinality function $|I_{ac}(\lambda)|_0$ changes by 1 at each breakpoint $\lambda_b\in\Lambda_b$ with probability 1.
2. If $n<d$, the cardinality function $|I_{ac}(\lambda)|_0$ changes by at most $\frac{(d-n)(d-n+1)}{2}+1$ at each breakpoint $\lambda_b\in\Lambda_b$ with probability 1.

To understand Theorem 1, notice that if $\lambda=0$, then $\text{supp}(S^{\text{opt}})$ is a complete graph almost surely. On the other hand, if λ is large enough, the edge set of $\text{supp}(S^{\text{opt}})$ is empty. The first implication of Theorem 1 is that whenever the number of available samples is close to d or higher, the

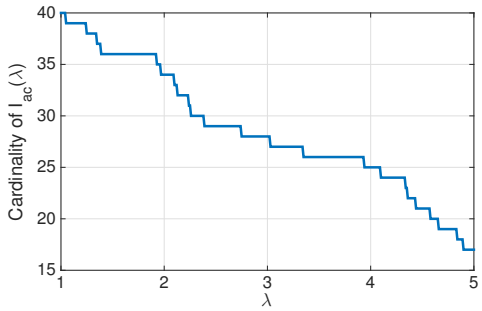


Fig. 1: The gradual change in the cardinality of $I_{ac}(\lambda)$ with respect to λ.

sparsity of supp(S^{opt}) changes gradually with respect to λ and there is no abrupt change in the edge set of supp(S^{opt}). The second implication applies to the scenario $n \geq d$. In this case, the first part of Theorem 1 unveils that, given an arbitrary integer value between 0 and $d(d-1)/2$, there is a nonempty interval for λ such that the number of edges in the corresponding sparsity graph supp(S^{opt}) is exactly equal to that number. This means that every level of sparsity can be achieved by Graphical Lasso via updating the regularization parameter λ appropriately (e.g., using a bisection technique). It also follows from the proof of the above theorem that each change in the sparsity pattern corresponds to the addition or removal of a single edge of the graph. The term "probability 1" in the above theorem means that the results hold if the sample covariance matrix is *generic*. This is always satisfied if the elements of Σ are allowed to be perturbed infinitesimally. Note that the genericity condition is met in practice due to the rounding errors of finite-precision machines.

*Example 1 (***randomly generated data):** In this example, assume that the true covariance matrix is equal to $NN^\top$, where N is a 10×10 randomly generated matrix whose entries are drawn from a normal distribution. We construct a sample covariance matrix using 10 independent samples (namely, $n = d = 10$). The behavior of the cardinality of $I_{ac}(\lambda)$ as a function of λ is illustrated in Figure 1. As expected from Theorem 1, the cardinality function $|I_{ac}(\lambda)|_0$ changes by one at each breakpoint.

In the remainder of this paper, it is desirable to show that Graphical Lasso has a closed-form solution if an acyclic support graph is sought, under mild conditions (an acyclic graph is an arbitrary graph with no cycles or loops). Consider the nonzero elements of the upper triangular part of Σ, excluding the diagonal elements. Let $\sigma_1, \sigma_2, ..., \sigma_{d(d-1)/2}$ denote the absolute values of those nonzero entries such that

$$\sigma_1 > \sigma_2 > ... > \sigma_{d(d-1)/2} > 0 \tag{5}$$

Note that $\sigma_{d(d-1)/2} \neq 0$ with probability 1 since Σ_* is non-singular and we have a finite number of samples (this is due to an implicit genericity assumption about Σ).

Definition 3: Consider an arbitrary positive regularization parameter λ that does not belong to the discrete set $\{\sigma_1, \sigma_2, ..., \sigma_{d(d-1)/2}\}$. Define the index k associated with λ as an integer number satisfying the relation $\lambda \in (\sigma_k, \sigma_{k+1})$. If λ is greater than σ_1, then k is set to zero.

Definition 4: Define the **residue of Σ at level k with respect to λ** as the matrix $\Sigma^{\text{res}}(k, \lambda)$ whose $(i,j)^{\text{th}}$ entry is equal to $\Sigma_{ij} - \lambda \times \text{sign}(\Sigma_{ij})$ if $|\Sigma_{ij}| > \lambda$ and is zero otherwise.

For notational simplicity, we use Σ^{res} instead of $\Sigma^{\text{res}}(k, \lambda)$ whenever the equivalence is implied by the context. Throughout this paper, the notation S^{opt} refers to the unique optimal solution of Graphical Lasso, rather than (1).

Definition 5: Define $T(S^{\text{opt}}, \lambda)$ as a $d \times d$ symmetric matrix whose $(i,j)^{\text{th}}$ entry is equal to $\Sigma_{ij} + \lambda \times \text{sign}(S^{\text{opt}}_{ij})$ if $(i,j) \in \text{supp}(S^{\text{opt}})$ and is zero otherwise.

The second main theorem of this paper is presented below.

Theorem 2: If the graph supp(S^{opt}) is acyclic and the matrix $I_d + T(S^{\text{opt}}, \lambda)$ is positive definite, then the following statements hold:

1. $\mathcal{E}^{\text{opt}} \subseteq \mathcal{E}^{\text{res}}$
2. The optimal solution S^{opt} of Graphical Lasso can be computed via the explicit formula

$$S^{\text{opt}}_{ij} = \begin{cases} 1 + \sum\limits_{(i,m) \in \mathcal{E}^{\text{opt}}} \frac{{\Sigma^{\text{res}}_{im}}^2}{1 - {\Sigma^{\text{res}}_{im}}^2} & \text{if} \quad i = j \\ \frac{-\Sigma^{\text{res}}_{ij}}{1 - {\Sigma^{\text{res}}_{ij}}^2} & \text{if} \quad (i,j) \in \mathcal{E}^{\text{opt}} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

   where $\mathcal{E}^{\text{opt}}$ and $\mathcal{E}^{\text{res}}$ denote the edge sets of supp(S^{opt}) and supp(Σ^{res}), respectively.

When the regularization parameter λ is large, the graph supp(S^{opt}) is expected to be sparse, and possibly acyclic. In this case, the matrix $T(S^{\text{opt}}, \lambda)$ is sparse with small nonzero entries. Under this circumstance, Theorem 2 reveals two important properties of the solution of Graphical Lasso: 1) its support graph is contained in the sparsity graph of the thresholded sample correlation matrix, and 2) the entries of this matrix can be found using the explicit formula (6). However, this formula requires to know the locations of the nonzero elements of S^{opt}. In what follows, we will replace the assumptions of the above theorem with easily verifiable rules that are independent from the optimal solution S^{opt} or the locations of its nonzero entries. Furthermore, it will be shown that these conditions are expected to hold when λ is large enough, i.e., if a sparse matrix S^{opt} is sought.

Theorem 3: Assume that the following conditions are satisfied:

C1. The graph supp(Σ^{res}) is acyclic.
C2. $I_d + \Sigma^{\text{res}}$ is positive definite.
C3. $(\sigma_1 - \lambda)^2 \leq \lambda - \sigma_{k+1}$.

Then, the sparsity pattern of the optimal solution S^{opt} of Graphical Lasso corresponds to the sparsity pattern of Σ^{res} and, in addition, S^{opt} can be obtained via the explicit formula (6).

The above theorem states that if a sparse graph is sought, then as long as some easy-to-verify conditions are met, there is an explicit formula for the optimal solution of Graphical Lasso. Condition C1 corresponds to the acyclic behavior

of the graph supp(S^{opt}) and is expected to be satisfied precisely or approximately if the regularization coefficient is sufficiently large (see [21] for more details). Condition C2 implies that the eigenvalues of the residue of Σ at level k with respect to λ should be greater than -1. This condition is also expected to be met for sparse graphs since most of the elements of Σ^{res} are equal to zero and the nonzero elements belong to $(-1+\lambda, 1-\lambda)$ (note that λ is large for sparse matrices). In particular, this condition is satisfied if $I_d + \Sigma^{\text{res}}$ is diagonally dominant. Moreover, with a natural choice of $\lambda = \frac{1}{2}(\sigma_k + \sigma_{k+1})$, Condition C3 can be written as

$$\left(2(\sigma_1 - \sigma_k) + (\sigma_k - \sigma_{k+1})\right)^2 \leq 2(\sigma_k - \sigma_{k+1}) \qquad (7)$$

Since $0 < \sigma_{k+1} < \sigma_k < \sigma_1 < 1$, both $\sigma_k - \sigma_{k+1}$ and $\sigma_1 - \sigma_k$ are in the interval $(0, 1)$ (the exclusion of 0 and 1 is due to the mild assumption that all off-diagonal nonzero elements of Σ are distinct). This leads to the next corollary.

Corollary 1: Define c as $\sigma_1 - \sigma_k$ and let λ be chosen as $(\sigma_k + \sigma_{k+1})/2$. Condition C3 is satisfied under either of the following conditions:

- $0 \leq c \leq 0.236$ and $(1-2c) - \sqrt{1-4c} \leq \sigma_k - \sigma_{k+1}$
- $0.236 \leq c \leq 0.25$ and $(1-2c) - \sqrt{1-4c} \leq \sigma_k - \sigma_{k+1} \leq (1-2c) + \sqrt{1-4c}$.

Furthermore, Condition C3 cannot be satisfied if $c > 0.25$. It can be inferred from the above corollary that if $\sigma_1 - \sigma_k$ is small enough (or if λ is large enough), Condition C3 is likely to be satisfied. In particular, if there is a clear separation between $\{\sigma_1, ..., \sigma_k\}$ (whose corresponding elements in Σ^{res} are nonzero) and $\{\sigma_{k+1}, ..., \sigma_{d(d-1)/2}\}$ (whose corresponding elements in Σ^{res} are set to zero), then Condition C3 is satisfied and one can use Theorem 3 to obtain the solution of Graphical Lasso without solving (2) numerically. Having computed the sample correlation matrix, we next show that checking the conditions in Theorem 3 and finding S^{opt} using (6) can all be carried out efficiently.

Corollary 2: Given Σ and λ, the total time complexity of checking the conditions in Theorem 3 and finding S^{opt} using (6) is $\mathcal{O}(d^2)$.

Theorems 2 and 3 significantly generalize the results of the recent paper [18]. That work gives sufficient conditions for the equivalence of the sparsity structures found using Graphical Lasso and the simple method of thresholding out the small entries of the sample correlation matrix. Those conditions imply that as long as the gap between the largest thresholded and smallest un-thresholded values in the sample correlation matrix is not too small, the matrix Σ after thresholding (by keeping only $2k$ off-diagonal entries of this matrix) has the same sparsity structure as S^{opt}. While the paper [18] sheds light on the relationship between Graphical Lasso and the thresholding technique, it suffers from two drawbacks: (1) the conditions guaranteeing this equivalence are not easy to verify, and (2) although it finds the correct sparsity structure of S^{opt}, i.e., the locations of its nonzero elements, it fails to obtain the exact values of its entries. In contrast, Theorem 3 offers a set of easily verifiable conditions for acyclic graphs, as well as a closed-form solution for Graphical Lasso in this case.

Remark 1: In this work, it is shown under some mild assumptions that the GL has an explicit closed-form solution if the support graph of the thresholded sample correlation matrix is acyclic. This approach can be adopted to find approximate solutions for the GL problem in the case where the support graph is not acyclic but is sparse. The main idea is to revise the closed-form formula and show that it approximately satisfies the optimality conditions for sparse graphs. Due to space restrictions, this generalization is not included in this paper and is only discussed in the technical report [21]. Furthermore, the results of [19] and [20] indicate that our closed form solution partially reveals the optimal solution of GL corresponding to disjoint and acyclic components of thresholded sample correlation matrix.

Next, we will illustrate the effectiveness of Theorem 3 in two examples.

*Example 2 (***synthetic data):** Suppose that the sample correlation matrix Σ has the following properties:

- Its diagonal elements are normalized to 1.
- The elements corresponding to an arbitrary spanning tree of supp(Σ) belong to the union of the intervals $[-0.85, -0.95]$ and $[0.85, 0.95]$.
- The off-diagonal entries that do not belong to the spanning tree are in the interval $[-0.85 + \omega, 0.85 - \omega]$.

The goal is to find conditions on λ, ω, and the size of the correlation matrix such that Theorem 3 can be used to obtain a closed-form solution for the GL problem. One can choose $\sigma_d < \lambda$ in order to ensure that supp(Σ^{res}) is acyclic. In particular, if we pick $\sigma_d < \lambda < \sigma_{d-1}$, the graph supp($\Sigma^{\text{res}}$) will be a spanning tree. Suppose that $\lambda = 0.85 - \epsilon$ for a sufficiently small number ϵ. First, consider Condition C2 in Theorem 3. One can easily verify that $I_d + \Sigma^{\text{res}}$ is positive definite if the inequality $\frac{1}{\deg(v)} > (\sigma_1 - \lambda)^2$ holds for every node v in supp(Σ^{res}), where deg(v) is the degree of node v. This condition is guaranteed to be satisfied for all acyclic graphs if $(d-1)(0.95-0.85)^2 < 1$ or $d \leq 100$. Regarding Condition C3, it can be observed that the relation $(\sigma_1 - \lambda)^2 \leq \lambda - \sigma_{k+1}$ holds if $(0.95 - 0.85)^2 < 0.85 - (0.85 - \omega)$. This implies that the inequality $\omega > 0.01$ guarantees the satisfaction of condition C3 for every acyclic graph supp(Σ^{res}). In other words, one can find the optimal solution of the GL problem explicitly using Theorem 3 as long as: 1) a spanning tree structure for the optimal solution of GL problem is sought, 2) the size of the correlation matrix is not greater than 100, and (3) the difference between σ_{d-1} and σ_d is greater than 0.01. Note that Condition (2) is conservative and can be removed for certain types of graphs (e.g., path graphs). The reason is that the matrix $I_d + \Sigma^{\text{res}}$ is always positive-definite for such graphs for all values of d.

*Example 3 (***electrical circuit):** In this example, we consider a resistor-capacitor (RC) circuit with 10 nodes. The goal is to estimate the structure of the underlying connectivity of the circuit based on the available noisy measurements of the nodal voltages. Here, we assume that the circuit elements are under a white Gaussian noise, known as Johnson-Nyquist noise. Recently, it has been shown in [6]

that the true inverse covariance matrix Σ_*^{-1} coincides with the capacitance matrix C of the circuit, which describes the underlying physical structure of the circuit. This means that an accurate estimation of Σ_*^{-1} would enable us to find C. Suppose that the true RC circuit has the structure illustrated in Figure 2a. Notice that the graph representing the structure of this circuit is a spanning tree. For simplicity, assume that the values of the resistors and capacitors are all equal to 1. The capacitance matrix C is defined as

$$C_{ij} = \begin{cases} -1 & \text{if } (i,j) \in \mathcal{E} \\ \alpha_{ii} - \sum_{(i,k)\in\mathcal{E}} C_{ik} & \text{if } i = j \end{cases} \quad (8)$$

where $\mathcal{E}$ is the edge set of the graph representation of the RC circuit and α_{ii} is the value of both the capacitor and the resistor connected to the ground at node i. Assume that α_{ii} is equal to 0.1 for $i = 5, 6$ and is zero for the remaining nodes. Let $V(t)$ denote the vector of voltages as a function of time t. In order to estimate C, one can use the sample covariance matrix $\frac{1}{n}\sum_{i=1}^{n} V(t_i)V(t_i)^\top$, where t_i is the i^{th} sampling time. The full representation of the normalized sample correlation matrix can be found in [21] and is omitted here due to space restrictions. Suppose that we know *a priori* that the graph capturing the structure of the circuit is a spanning tree. Based on the sample correlation matrix, we have $\sigma_1 = 0.9406$, $\sigma_9 = 0.9000$, and $\sigma_{10} = 0.8966$. One can easily verify that if we set $\lambda = 0.9000 - \epsilon$ for a sufficiently small $\epsilon > 0$, the graph $\text{supp}(\Sigma^{\text{res}})$ is a spanning tree that satisfies Condition C1 in Theorem 3. Furthermore, this choice of λ ensures the positive definiteness of $I_d + \Sigma^{\text{res}}$, which guarantees the satisfaction of Condition C2. Moreover, note that $(\sigma_1 - \lambda)^2 = (0.0406 + \epsilon)^2$ and $\lambda - \sigma_{10} = 0.0033 - \epsilon$. This implies that Condition C3 is guaranteed to be satisfied for small values of ϵ. Therefore, Theorem 3 can be used to find the closed-form solution of the GL problem. The optimal solution has the sparsity structure depicted in Figure 2b. It can be observed that the estimated sparsity structure coincides with the true physical structure of the RC circuit.

III. Proofs

In this section, the proofs of the results presented before are provided. A number of lemmas are required for this purpose.

Lemma 1: The space of feasible sample covariance matrices Σ has dimension $nd - \frac{n(n-1)}{2}$ if $n < d$ and $\frac{d(d+1)}{2}$ if $n \geq d$.

Proof: The proof is omitted for brevity. The details can be found in [21]. ∎

Before presenting the proof of Theorem 1, it is desirable to study the behavior of $S^{\text{opt}}(\lambda)$ and the set of breakpoints Λ_b.

Lemma 2: The minimizer $S^{\text{opt}}(\lambda)$ is continuous in λ.

Proof: The proof follows immediately from the result in [23]. ∎

Corollary 3: Λ_b is a countable set.

Proof: It is straightforward to verify that the continuity of $S^{\text{opt}}(\lambda)$ (due to Lemma 2) implies that Λ_b is a countable set. ∎

The unique solution of the GL problem can be characterized based on the KKT conditions given below.

Lemma 3 ([17]): The matrix S^{opt} is the optimal solution of Graphical Lasso if and only if it satisfies the following conditions for all $i, j \in \{1, 2, ..., d\}$:

$$(S^{\text{opt}})_{ij}^{-1} = \Sigma_{ij} \quad \text{if} \quad i = j \quad (9a)$$
$$(S^{\text{opt}})_{ij}^{-1} = \Sigma_{ij} + \lambda \times \text{sign}(S_{ij}^{\text{opt}}) \quad \text{if} \quad S_{ij}^{\text{opt}} \neq 0 \quad (9b)$$
$$\Sigma_{ij} - \lambda \leq (S^{\text{opt}})_{ij}^{-1} \leq \Sigma_{ij} + \lambda \quad \text{if} \quad S_{ij}^{\text{opt}} = 0 \quad (9c)$$

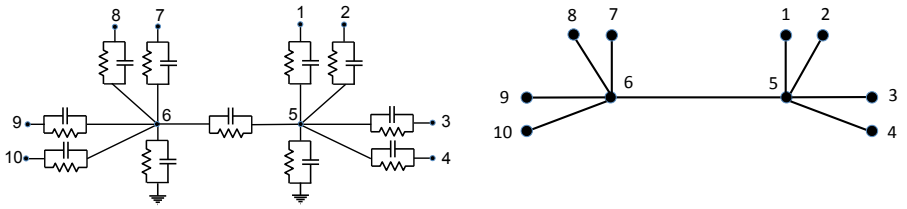
where $(S^{\text{opt}})_{ij}^{-1}$ denotes the $(i,j)^{\text{th}}$ entry of $(S^{\text{opt}})^{-1}$.

Using the aforementioned lemmas, we are ready to present the proof of Theorem 1.

Proof of Theorem 1: We prove the first part of the theorem below. The proof of the second part of the theorem is omitted since it follows by adopting a similar argument. To simplify the presentation, assume that Σ is a sample covariance matrix, rather than a sample correlation matrix (this implies that the diagonal entries are not necessarily normalized to 1).

By contradiction, suppose that $|I_{ac}(\lambda)|_0$ does not change by 1 as λ passes through a breakpoint λ_b. There are three possibilities: (i) $|I_{ac}(\lambda)|_0$ increases by at least 2, (ii) $|I_{ac}(\lambda)|_0$ decreases by at least 2, (iii) the cardinality remains the same but at least one new element enters $I_{ac}(\lambda)$ and exactly the same number of elements leave $I_{ac}(\lambda)$. With no loss of generality, we investigate only scenario (i) in this proof under the assumption that exactly two elements are added to the set $I_{ac}(\lambda)$ without any element leaving the set. First, we show that as λ approaches λ_b from both sides, the sample covariance matrix must satisfy a particular set of equations. Then, we prove that these equations are satisfied in the case $n \geq d$ with probability zero or for a generic Gaussian random vector with a non-singular sample covariance matrix (genericity is achieved by an infinitesimal perturbation of the entries of the covariance matrix).

Assume that λ_b^+ is the smallest breakpoint that is greater than λ_b. Similarly, let λ_b^- denote the largest breakpoint that is less than λ_b. It may occur that either λ_b^+ or λ_b^- does not exist if λ_b is the smallest/largest breakpoint in Λ_b, in which case λ_b^- and λ_b^+ can be set to 0 and $+\infty$, respectively (note that 0 does not belong to Λ_b since Σ does not have zero entries at $\lambda = 0$ with probability 1). Because Λ_b is a discrete set in light of Corollary 3, one can write $\lambda_b^- < \lambda_b < \lambda_b^+$. Consider a number $r \geq d$ such that $|I_{ac}(\lambda_b + \epsilon^+)|_0 = r$ for every $0 < \epsilon^+ < \lambda_b^+ - \lambda_b$. For simplicity, we index the elements of S and S^{-1} with natural numbers rather than pairs of numbers based on the vectorized versions of these matrices. Denote the active set $I_{ac}(\lambda_b + \epsilon^+)$ of $S^{\text{opt}}(\lambda_b + \epsilon^+)$ as $\{s_1^*(\lambda_b + \epsilon^+), s_2^*(\lambda_b + \epsilon^+), ..., s_r^*(\lambda_b + \epsilon^+)\}$, where the first d elements correspond to the diagonal entries of S^{opt} and the remaining elements correspond to its nonzero off-diagonal upper-triangular entries. For every $i \notin I_{ac}(\lambda_b + \epsilon^+)$, we have $s_i^*(\lambda_b + \epsilon^+) = 0$. From the optimality conditions in Lemma



(a) RC circuit (b) Estimated sparsity graph of RC circuit

Fig. 2: a) The physical structure of the RC circuit with 10 nodes considered in Example 3. b) The sparsity graph of the optimal solution of the GL problem in Example 3.

3, one can write:

$$(S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b+\epsilon^+)} = \Sigma_i \quad \text{if} \quad 1 \le i \le d \tag{10a}$$

$$(S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b+\epsilon^+)} = \Sigma_i + (\lambda_b+\epsilon^+) \times \text{sign}(S_i|_{S=S^{\text{opt}}(\lambda_b+\epsilon^+)}) \quad \text{if} \quad d+1 \le i \le r \tag{10b}$$

$$\Sigma_i - (\lambda_b+\epsilon^+) \le (S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b+\epsilon^+)} \le \Sigma_i + \lambda_b + \epsilon^+ \quad \text{if} \quad i > r \tag{10c}$$

By assumption, the relation $|I_{ac}(\lambda_b - \epsilon^-)|_0 = r+2$ holds for every $0 < \epsilon^- < \lambda_b - \lambda_b^-$. This means that as λ decreases to pass through λ_b, two elements of S^{opt} will be added to the set of the nonzero elements. Denote these new elements as s^*_{r+1} and s^*_{r+2}. Then, the optimality conditions at $\lambda_b - \epsilon^-$ can be written as

$$(S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b-\epsilon^-)} = \Sigma_i \quad \text{if} \quad 1 \le i \le d \tag{11a}$$

$$(S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b-\epsilon^-)} = \Sigma_i + (\lambda_b-\epsilon^-) \times \text{sign}(S_i|_{S=S^{\text{opt}}(\lambda_b-\epsilon^-)}) \quad \text{if} \quad d+1 \le i \le r \tag{11b}$$

$$\Sigma_i - (\lambda_b-\epsilon^-) \le (S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b-\epsilon^-)} \le \Sigma_i + \lambda_b - \epsilon^- \quad \text{if} \quad i > r \tag{11c}$$

Without loss of generality in the analysis, we drop the sign functions from (10) and (11) for now. Consider the limiting behavior of $s^*_{r+1}(\lambda)$ and $s^*_{r+2}(\lambda)$. Note that $s^*_{r+1}(\lambda_b+\epsilon^+) = s^*_{r+2}(\lambda_b+\epsilon^+) = 0$ for every $0 < \epsilon^+ < \lambda_b^+ - \lambda_b$. Due to the continuity of $S^{\text{opt}}(\lambda)$ (as shown in Lemma 2), one can write:

$$s^*_{r+1}(\lambda_b) = \lim_{\epsilon^+ \to 0^+} s^*_{r+1}(\lambda_b+\epsilon^+) = 0 \tag{12a}$$

$$s^*_{r+2}(\lambda_b) = \lim_{\epsilon^+ \to 0^+} s^*_{r+2}(\lambda_b+\epsilon^+) = 0 \tag{12b}$$

Notice that since $\Sigma \succ 0$, we obtain $S^{\text{opt}}(\lambda) \succ 0$ for every finite number λ. Therefore, $(S^{\text{opt}}(\lambda))^{-1}$ is well-defined and a continuously differentiable function of λ. It follows from (11) that

$$(S)_{r+1}^{-1}|_{S=S^{\text{opt}}(\lambda_b)} = \lim_{\epsilon^- \to 0^+} (S)_{r+1}^{-1}|_{S=S^{\text{opt}}(\lambda_b-\epsilon^-)} = \Sigma_{r+1} + \lambda_b \tag{13a}$$

$$(S)_{r+2}^{-1}|_{S=S^{\text{opt}}(\lambda_b)} = \lim_{\epsilon^- \to 0^+} (S)_{r+2}^{-1}|_{S=S^{\text{opt}}(\lambda_b-\epsilon^-)} = \Sigma_{r+2} + \lambda_b \tag{13b}$$

Thus,

$$(S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b)} = \lim_{\epsilon^- \to 0^+} (S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b-\epsilon^-)} = \Sigma_i, \quad \forall i \in \{1...,d\} \tag{14a}$$

$$(S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b)} - \lambda_b = \lim_{\epsilon^- \to 0^+} (S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b-\epsilon^-)} - \lambda_b = \Sigma_i, \quad \forall i \in \{d+1...,r+2\} \tag{14b}$$

Now, notice that $(S)_i^{-1}|_{S=S^{\text{opt}}(\lambda_b)}$ can be written as a function of the nonzero elements of S^{opt}, i.e. $\{s_1^*(\lambda_b),...,s_r^*(\lambda_b)\}$. Let the left-hand side in (14a) and (14b) be defined as a function denoted by $f_i(s_1^*,...,s_r^*,\lambda_b)$ for $i = 1,2,...,r+2$ (for simplicity, we dropped λ from s_i^*). Therefore, the set $\{s_1^*,...,s_r^*,\lambda_b\}$ must satisfy the following set of equations:

$$f_i(s_1^*,...,s_r^*,\lambda_b) = \Sigma_i, \qquad \forall i \in \{1,2,...,r+2\} \tag{15}$$

In light of (15), there exist $r+2$ functions, each with $r+1$ arguments, that must be equal to the given numbers $\{\Sigma_1,...,\Sigma_{r+2}\}$ at some point. Notice that the image of this set of functions has dimension of at most $r+1$ (due to the dimension of its domain). Therefore, $\{\Sigma_1,...,\Sigma_{r+2}\}$ should belong to a $(r+1)$–dimensional manifold. This implies that Σ should belong to a $((d^2+d)/2-1)$–dimensional manifold. One can verify that we may have up to a finite number of distinct active sets for all values of λ. Furthermore, recall that the sign operator was dropped throughout the proof. For each setting of the sign operator, one can show that the same argument can be made and a set of equations similar to (15) should be satisfied. Therefore, $|I_{ac}(\lambda)|_0$ may change by at least 2 only if Σ resides in a finite union of $((d^2+d)/2-1)$-dimensional manifolds. This occurs with probability zero since the set of feasible sample covariance matrices has dimension $(d^2+d)/2$ due to Lemma 1. ■

Definition 6: The complement of the graph $\mathcal{G}$ is defined as a graph $\mathcal{G}^{(c)}$ with the same vertex set as $\mathcal{G}$ whose edge set is the complement of the edge set of $\mathcal{G}$.

Definition 7: Given two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with the same vertex set, the graph $\mathcal{G}_1$ is called a subgraph of $\mathcal{G}_2$ if its edge set is a subset of the edge set of $\mathcal{G}_2$. This inclusion is shown as $\mathcal{G}_1 \subseteq \mathcal{G}_2$.

Definition 8: A symmetric matrix M is **inverse-consistent** if there exists another matrix N with zero

diagonal entries such that

$$M + N \succ 0 \tag{16a}$$
$$\text{supp}(N) \subseteq (\text{supp}(M))^{(c)} \tag{16b}$$
$$\text{supp}\left((M+N)^{-1}\right) \subseteq \text{supp}(M) \tag{16c}$$

A matrix N satisfying the above properties is called **inverse-consistent complement** and is denoted by $M^{(c)}$. Moreover, M is called **sign-consistent** if the $(i,j)^{\text{th}}$ entries of M and $(M + M^{(c)})^{-1}$ have opposite signs for every $(i,j) \in \text{supp}(M)$.

Definition 9: Given a graph $\mathcal{G}$ and a positive number α, the function $\beta(\mathcal{G},\alpha)$ is defined as the maximum of $\|M^{(c)}\|_{\max}$ over all inverse-consistence positive-definite matrices M with diagonal entries equal to 1 such that $\text{supp}(M) \subseteq \mathcal{G}$ and $\|M\|_{\max} \leq \alpha$.

Lemma 4 ([18]): The following statements hold:

- Every positive definite matrix is inverse-consistent and has a unique inverse-consistent complement.
- The support graphs of S^{opt} and $\Sigma^{\text{res}}(k,\lambda)$ are equivalent if three conditions hold: (i) $I_d + \Sigma^{\text{res}}$ is positive definite, (ii) $I_d + \Sigma^{\text{res}}$ is sign-consistent, (iii) the inequality $\beta(\text{supp}(\Sigma^{\text{res}}), \sigma_1 - \lambda) \leq \lambda - \sigma_{k+1}$ is satisfied. Moreover, Condition (ii) is implied by Condition (i) if $\text{supp}(\Sigma^{\text{res}})$ is acyclic.

To be able to use the above lemma, we need to analytically calculate the function $\beta(\mathcal{G},\alpha)$. This will be addressed next.

Lemma 5: The relation $\beta(\mathcal{G},\alpha) \leq \alpha^2$ holds for every $0 \leq \alpha < 1$ if $\mathcal{G}$ is acyclic. Furthermore, strict equality holds if $\mathcal{G}$ includes a path of length at least 2.

Proof:

Without loss of generality, assume that $\mathcal{G}$ is a tree (if there are disjoint components, the argument made in the sequel can be used for every connected component of $\mathcal{G}$). Let $d_{i,j}$ denote the unique path between every two disparate nodes i and j in $\mathcal{G}$. Furthermore, define $N(i)$ as the set of all neighbors of node i in $\mathcal{G}$. Consider a positive definite matrix M with diagonal elements equal to 1 such that $\|M\|_{\max} = \alpha$ and $\text{supp}(M) = \mathcal{G}$. Define the matrix N as

$$N_{ij} = \begin{cases} \prod_{(m,t)\in d_{i,j}} M_{mt} & \text{if} \quad (i,j) \in (supp(M))^{(c)} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Moreover, define

$$A_{ij} = \begin{cases} 1 + \sum_{m\in N(i)} \frac{M_{mi}^2}{1-M_{mi}^2} & \text{if} \quad i = j \\ \frac{-M_{ij}}{1-M_{ij}^2} & \text{if} \quad (i,j) \in supp(M) \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

The goal is to show that the matrix N is the unique inverse-consistent complement of M. First, note that $\text{supp}(N) = (\text{supp}(M))^{(c)}$ and $\text{supp}(M) = \text{supp}(A)$. Next, it is desirable to prove that $(M+N)^{-1} = A$ or equivalently $(M+N)A = I$. Upon defining $T = (M+N)A$, one can write:

$$\begin{aligned} T_{ii} &= \sum_{m=1}^{d} (M_{im} + N_{im}) A_{mi} \\ &= 1 + \sum_{m\in N(i)} \frac{M_{mi}^2}{1-M_{mi}^2} - \sum_{m\in N(i)} \frac{M_{mi}^2}{1-M_{mi}^2} = 1 \end{aligned} \tag{19}$$

Moreover, for every pair of nodes i and j, define D_{ij} to be equal to $\prod_{(k,t)\in d_{i,j}} M_{kt}$ if $i \neq j$ and equal to 1 if $i = j$. Consider a pair of distinct nodes i and j. Let t denote the node adjacent to j in $d_{i,j}$ (note that we may have $t = i$). It can be verified that

$$\begin{aligned} T_{ij} = \sum_{m=1}^{d} (M_{im} + N_{im}) A_{mj} &= D_{ij} \left(1 + \sum_{m\in N(j)} \frac{M_{mj}^2}{1-M_{mj}^2}\right) \\ &- D_{it} \left(\frac{M_{tj}}{1-M_{tj}^2}\right) - \sum_{\substack{m\in N(j) \\ m\neq t}} D_{im} \frac{M_{mj}}{1-M_{mj}^2} \end{aligned} \tag{20}$$

Furthermore,

$$\begin{aligned} &D_{ij} = D_{it} M_{tj}, \\ &D_{im} = D_{it} M_{tj} M_{jm}, \qquad \forall\ m \in N(j),\ m \neq t \end{aligned} \tag{21}$$

Plugging (21) into (20) yields that

$$\begin{aligned} T_{ij} =& D_{it} M_{tj} \left(\frac{1}{1-M_{tj}^2} + \sum_{\substack{m\in N(j) \\ m\neq t}} \frac{M_{mj}^2}{1-M_{mj}^2}\right) - D_{it} \left(\frac{M_{tj}}{1-M_{tj}^2}\right) \\ &- D_{it} M_{tj} \sum_{\substack{m\in N(j) \\ m\neq t}} \frac{M_{mj}^2}{1-M_{mj}^2} = 0 \end{aligned} \tag{22}$$

Hence, $T = I$. Finally, we need to show that $M + N \succ 0$. To this end, it suffices to prove that $A \succ 0$. Note that A can be written as $I + \sum_{(i,j)\in\mathcal{G}} L^{(i,j)}$, where $L^{(i,j)}$ is defined as

$$L_{rl}^{(i,j)} = \begin{cases} \frac{M_{ij}^2}{1-M_{ij}^2} & \text{if} \quad r = l = i \text{ or } j \\ \frac{-M_{ij}}{1-M_{ij}^2} & \text{if} \quad (r,l) = (i,j) \\ 0 & \text{otherwise} \end{cases}$$

Consider the term $x^T A x$ for an arbitrary vector $x \in \mathbb{R}^d$. One can verify that

$$\begin{aligned} x^T A x &= \sum_{i=1}^{d} x_i^2 + \sum_{(i,j)\in\mathcal{G}} x^T L^{(i,j)} x \\ &= \sum_{i=1}^{d} x_i^2 + \sum_{(i,j)\in\mathcal{G}} \left(\frac{M_{ij}^2}{1-M_{ij}^2}\right) x_i^2 + \left(\frac{M_{ij}^2}{1-M_{ij}^2}\right) x_j^2 \\ &\quad - \left(\frac{2M_{ij}}{1-M_{ij}^2}\right) x_i x_j \end{aligned} \tag{23}$$

Without loss of generality, assume that the graph is a rooted tree with the root at node d. Assume that each edge (i,j)

defines a direction that is toward the root. Then, it follows from (23) that

$$\begin{aligned} x^T A x =& x_d^2 + \sum_{(i,j)\in\mathcal{G}} \left(\frac{1}{1-M_{ij}^2}\right) x_i^2 + \left(\frac{M_{ij}^2}{1-M_{ij}^2}\right) x_j^2 \\ & - \left(\frac{2M_{ij}}{1-M_{ij}^2}\right) x_i x_j \\ =& x_d^2 + \sum_{(i,j)\in\mathcal{G}} \frac{(x_i - M_{ij}x_j)^2}{1-M_{ij}^2} \geq 0 \end{aligned} \tag{24}$$

Therefore, $M+N \succeq 0$ and because it is invertible, we have $M+N \succ 0$. Hence, according to Definition 8 and Lemma 4, the matrix N is the unique inverse-consistent compliment of M. On the other hand, due to the definition of N, we have $\|N\|_{\max} \leq \alpha^2$ and consequently $\beta(\mathcal{G},\alpha) \leq \alpha^2$. Now, suppose that $\mathcal{G}$ includes a path of length at least 2, e.g., the edges $(1,2)$ and $(2,3)$ belong to $\mathcal{G}$. By setting $M_{12} = M_{23} = \alpha$ and choosing sufficiently small values for those entries of M corresponding to the remaining edges in $\mathcal{G}$, the matrix M remains positive definite and we obtain $\|N\|_{\max} = \alpha^2$. This completes the proof. ■

Proof of Theorem 3: Based on Lemmas 4 and 5, the conditions introduced in Part 2 of Lemma 4 can be reduced to Conditions C2 and C3 in Theorem 3 if $\text{supp}(\Sigma^{\text{res}})$ is acyclic. Moreover, suppose that M is set to $I_d + \Sigma^{\text{res}}$, and that the matrices N and A are defined as (17) and (18), respectively. It can be verified that $S^{\text{opt}} = A$ satisfies the optimality conditions given in (9). Therefore, it is the unique solution of the GL problem. ■

The proofs of Theorem 2 and Corollaries 1 and 2 are omitted for brevity and can be found in [21].

IV. Conclusion

This paper is concerned with the Graphical Lasso method, which is used to determine the conditional dependence between different elements in a multivariate Gaussian distribution using a limited number of samples. It is well-known that the sparsity level of the solution of Graphical Lasso heavily depends on the regularization coefficient. It is important to understand the behavior of the optimal solution of Graphical Lasso as a function of the regularization coefficient. To this end, it is shown in this paper that if the number of available samples is not too small, the sparsity pattern of the optimal solution of Graphical Lasso changes gradually. Based on this result, it is then proved that every level of sparsity is achievable in the solution of Graphical Lasso after fine-tuning the regularization coefficient. Although Graphical Lasso as an optimization problem is computationally prohibitive for large-scale problems, it is shown that this problem benefits from a simple closed-form solution for acyclic graphs. This explicit formula can be used to find approximate solutions for non-acyclic sparse graphs. The efficacy of the developed results is demonstrated on synthetic data and electrical circuits.

References

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[2] S. J. Benson, Y. Ye, and X. Zhang, "Solving large-scale sparse semidefinite programs for combinatorial optimization," *SIAM Journal on Optimization*, vol. 10, no. 2, pp. 443–461, 2000.

[3] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[4] 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, 2009.

[5] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, 2010.

[6] S. Sojoudi and J. Doyle, "Study of the brain functional network using synthetic data," *52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 350–357, 2014.

[7] M. Fardad, F. Lin, and M. R. Jovanović, "Sparsity-promoting optimal control for a class of distributed systems," *American Control Conference*, pp. 2050–2055, 2011.

[8] S. Fattahi and J. Lavaei, "On the convexity of optimal decentralized control problem and sparsity path," *American Control Conference*, 2017.

[9] Y. Zheng, R. P. Mason, and A. Papachristodoulou, "Scalable design of structured controllers using chordal decomposition," *IEEE Transactions on Automatic Control*, 2017.

[10] F. Lian, A. Chakrabortty, and A. Duel-Hallen, "Game-theoretic multi-agent control and network cost allocation under communication constraints," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 2, pp. 330–340, 2017.

[11] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

[12] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Basel: Birkhäuser, 2013, vol. 1, no. 3.

[13] P. Bühlmann and S. V. D. Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[14] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, pp. 101–148, 2010.

[15] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *Journal of Machine learning research*, vol. 9, pp. 485–516, 2008.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[17] S. Sojoudi, "Equivalence of graphical lasso and thresholding for sparse graphs," *Journal of Machine Learning Research*, vol. 17, no. 115, pp. 1–21, 2016.

[18] S. Sojoudi, "Graphical lasso and thresholding: Conditions for equivalence," *IEEE 55th Conference on Decision and Control*, pp. 7042–7048, 2016.

[19] R. Mazumder and T. Hastie, "Exact covariance thresholding into connected components for large-scale graphical lasso," *Journal of Machine Learning Research*, vol. 13, pp. 781–794, 2012.

[20] D. M. Witten, J. H. Friedman, and N. Simon, "New insights and faster computations for the graphical lasso," *Journal of Computational and Graphical Statistics*, vol. 20, no. 4, pp. 892–900, 2011.

[21] S. Fattahi and S. Sojoudi, "Graphical lasso and thresholding: Equivalence and closed-form solutions," *https://arxiv.org/abs/1708.09479*, 2017.

[22] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[23] H. Zhou and Y. Wu, "A generic path algorithm for regularized statistical estimation," *Journal of the American Statistical Association*, vol. 109, no. 506, pp. 686–699, 2014.