# Sparse-Group Non-convex Penalized Multi-Attribute Graphical Model Selection 

Jitendra K. Tugnait<br>Dept. of Electrical \& Computer Eng.<br>Auburn University, Auburn, AL 36849, USA


#### Abstract

We consider the problem of inferring the conditional independence graph (CIG) of high-dimensional Gaussian vectors from multi-attribute data. Most existing methods for graph estimation are based on single-attribute models where one associates a scalar random variable with each node. In multi-attribute graphical models, each node represents a random vector. In this paper we consider a sparse-group smoothly clipped absolute deviation (SG-SCAD) penalty instead of sparse-group lasso (SGL) penalty to regularize the problem. We analyze an SG-SCAD-penalized log-likelihood based objective function to establish consistency of a local estimator of inverse covariance. A numerical example is presented to illustrate the advantage of SG-SCAD-penalty over the usual SGL-penalty.


Keywords: Multi-attribute graph learning; inverse covariance estimation; undirected graph; SCAD penalty.

## I. Introduction

Graphical models provide a powerful tool for analyzing multivariate data [1], [2]. In an undirected graphical model, the conditional dependency structure among $p$ random variables $x_{1}, x_{1}, \cdots, x_{p},\left(\boldsymbol{x}=\left[\begin{array}{llll}x_{1} & x_{2} & \cdots & x_{p}\end{array}\right]^{\top}\right)$, is represented using an undirected graph $\mathcal{G}=(V, \mathcal{E})$, where $V=\{1,2, \cdots, p\}=$ $[p]$ is the set of $p$ nodes corresponding to the $p$ random variables $x_{i}$ 's, and $\mathcal{E} \subseteq[p] \times[p]$ is the set of undirected edges describing conditional dependencies among $x_{i}$ 's. The graph $\mathcal{G}$ then is a conditional independence graph (CIG) where there is no edge between nodes $i$ and $j$ iff $x_{i}$ and $x_{j}$ are conditionally independent given the remaining $p-2$ variables.

Gaussian graphical models (GGMs) are CIGs where $\boldsymbol{x}$ is multivariate Gaussian. Suppose $\boldsymbol{x}$ has positive-definite covariance matrix $\boldsymbol{\Sigma}$ with inverse covariance matrix $\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}$. Then $\Omega_{i j}$, the $(i, j)$-th element of $\Omega$, is zero iff $x_{i}$ and $x_{j}$ are conditionally independent. Given $n$ samples of $\boldsymbol{x}$, in highdimensional settings, one estimates $\Omega$ under some sparsity constraints; see [3]-[7]. In these graphs each node represents a scalar random variable. In many applications, there may be more than one random variable associated with a node. This class of graphical models has been called multi-attribute graphical models in [8]-[11]. Image graphs for color images with three variables (RGB) per pixel node, is an example of multi-attribute graphical models. Methods of [12], [13] concerned with grayscale images do not apply to color images. Recently in [14], a sparse-group lasso (SGL) based penalized log-likelihood approach for graph learning from multi-attribute

[^0]data was presented; in comparison, [9], [10] consider only group lasso, and therefore, are a special case of SGL.

Contributions: In this paper we consider a sparse-group smoothly clipped absolute deviation (SG-SCAD) penalty instead of SGL penalty [14], following group SCAD penalty [15]. The SCAD penalty was first exploited for graphical model selection in [16]. SCAD penalty can produce sparse set of solution like lasso, and approximately unbiased coefficients for large coefficients, unlike lasso. But this penalty is nonconvex, unlike lasso. Sufficient conditions for consistency of inverse covariance estimator and specification of its rate of convergence are provided in this paper. Such aspects are not considered in [15]; [11] deals with low-dimensional models.

Notation: Given $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, we use $\phi_{\min }(\boldsymbol{A}), \phi_{\max }(\boldsymbol{A})$, $|\mathbf{A}|$ and $\operatorname{tr}(\boldsymbol{A})$ to denote the minimum eigenvalue, maximum eigenvalue, determinant and trace of $\boldsymbol{A}$, respectively. For $\boldsymbol{B} \in$ $\mathbb{R}^{p \times q}$, we have $\|\boldsymbol{B}\|=\sqrt{\phi_{\max }\left(\boldsymbol{B}^{\top} \boldsymbol{B}\right)},\|\boldsymbol{B}\|_{F}=\sqrt{\operatorname{tr}\left(\boldsymbol{B}^{\top} \boldsymbol{B}\right)}$ and $\|\boldsymbol{B}\|_{1}=\sum_{i, j}\left|B_{i j}\right|$ where $B_{i j}$ is the $(i, j)$-th element of $\boldsymbol{B}$ (also denoted by $[\boldsymbol{B}]_{i j}$ ). Given $\boldsymbol{A} \in \mathbb{R}^{p \times p}, \boldsymbol{A}^{+}=\operatorname{diag}(\boldsymbol{A})$ is a diagonal matrix with the same diagonal as $\boldsymbol{A}$, and $\boldsymbol{A}^{-}=\boldsymbol{A}-\boldsymbol{A}^{+}$is $\boldsymbol{A}$ with all its diagonal elements set to zero. The notation $\boldsymbol{y}_{n}=\mathcal{O}_{P}\left(\boldsymbol{x}_{n}\right)$ for random vectors $\boldsymbol{y}_{n}, \boldsymbol{x}_{n} \in \mathbb{R}^{p}$ means that for any $\varepsilon>0$, there exists $0<M<\infty$ such that $P\left(\left\|\boldsymbol{y}_{n}\right\| \leq M\left\|\boldsymbol{x}_{n}\right\|\right) \geq 1-\varepsilon \forall n \geq 1$.

## II. System Model

We will call $\mathcal{G}$ considered in Sec. I a single-attribute graphical model for $\boldsymbol{x}$. Now consider $p$ jointly Gaussian random vectors $\boldsymbol{z}_{i} \in \mathbb{R}^{m}, i=1,2, \cdots, p$. We associate $\boldsymbol{z}_{i}$ with the $i$ th node of an undirected graph $\mathcal{G}=(V, \mathcal{E})$ where $V=[p]$ and edges in $\mathcal{E}$ describe the conditional dependencies among vectors $\left\{\boldsymbol{z}_{i}, i \in V\right\}$. As in the scalar case ( $m=1$ ), there is no edge between node $i$ and node $j$ in $\mathcal{G}$ iff random vectors $\boldsymbol{z}_{i}$ and $\boldsymbol{z}_{j}$ are conditionally independent given all the remaining random vectors [9], [10]. This is the multi-attribute Gaussian graphical model of interest in this paper.

Define the $m p$-vector

$$
\boldsymbol{x}=\left[\begin{array}{llll}
\boldsymbol{z}_{1}^{\top} & \boldsymbol{z}_{2}^{\top} & \cdots & \boldsymbol{z}_{p}^{\top} \tag{1}
\end{array}\right]^{\top} \in \mathbb{R}^{m p}
$$

Suppose we have $n$ i.i.d. observations $\boldsymbol{x}(t), t=0,1, \cdots, n-1$, of zero-mean $\boldsymbol{x}$. Our objective is to estimate the inverse covariance matrix $\left(\mathbb{E}\left\{\mathbf{x x}^{\top}\right\}\right)^{-1}$ and to determine if edge $\{i, j\} \in \mathcal{E}$, given data $\{\boldsymbol{x}(t)\}_{t=0}^{n-1}$. Let us associate $\boldsymbol{x}$ with an "enlarged" graph $\overline{\mathcal{G}}=(\bar{V}, \overline{\mathcal{E}})$, where $\bar{V}=[1, m p]$ and
$\overline{\mathcal{E}} \subseteq \bar{V} \times \bar{V}$. Now $\left[\boldsymbol{z}_{j}\right]_{\ell}$, the $\ell$ th component of $\boldsymbol{z}_{j}$ associated with node $j$ of $\mathcal{G}=(V, \mathcal{E})$, is the random variable $x_{q}=[\boldsymbol{x}]_{q}$, where $q=(j-1) m+\ell, j=1,2, \cdots, p$ and $\ell=1,2, \cdots, m$. The random variable $x_{q}$ is associated with node $q$ of $\overline{\mathcal{G}}=(\bar{V}, \overline{\mathcal{E}})$. Corresponding to the edge $\{j, k\} \in \mathcal{E}$ in the multi-attribute $\mathcal{G}=(V, \mathcal{E})$, there are $m^{2}$ edges $\{q, r\} \in \overline{\mathcal{E}}$ specified by $q=(j-1) m+s$ and $r=(k-1) m+t$, where $s=1,2, \cdots, m$ and $t=1,2, \cdots, m$. The graph $\overline{\mathcal{G}}=(\bar{V}, \overline{\mathcal{E}})$ is a single-attribute graph. In order for $\overline{\mathcal{G}}$ to reflect the conditional independencies encoded in $\mathcal{G}$, we must have the equivalence $\{j, k\} \notin \mathcal{E} \Leftrightarrow \overline{\mathcal{E}}^{(j k)} \cap \overline{\mathcal{E}}=\emptyset$, where $\overline{\mathcal{E}}^{(j k)}=\{\{q, r\} \mid q=$ $(j-1) m+s, r=(k-1) m+t, s, t=1,2, \cdots, m\}$. Let $\boldsymbol{R}_{x x}=\mathbb{E}\left\{\boldsymbol{x} \boldsymbol{x}^{\top}\right\} \succ \mathbf{0}$ and $\boldsymbol{\Omega}=\boldsymbol{R}_{x x}^{-1}$. Define the $(j, k)$ th $m \times m$ subblock $\boldsymbol{\Omega}^{(j k)}$ of $\boldsymbol{\Omega}$ as

$$
\begin{equation*}
\left[\boldsymbol{\Omega}^{(j k)}\right]_{s t}=[\boldsymbol{\Omega}]_{(j-1) m+s,(k-1) m+t}, s, t=1,2, \cdots, m \tag{2}
\end{equation*}
$$

It is established in [10, Sec. 2.1] that $\boldsymbol{\Omega}^{(j k)}=\mathbf{0} \Leftrightarrow\{j, k\} \notin \mathcal{E}$. Since $\boldsymbol{\Omega}^{(j k)}=\mathbf{0}$ is equivalent to $[\boldsymbol{\Omega}]_{q r}=0$ for every $\{q, r\} \in$ $\overline{\mathcal{E}}^{(j k)}$, and since, by [1, Proposition 5.2], $[\boldsymbol{\Omega}]_{q r}=0$ iff $x_{q}$ and $x_{r}$ are conditionally independent, hence, iff $\{q, r\} \notin \overline{\mathcal{E}}$, it follows that the aforementioned equivalence holds true.

## A. SG-SCAD-Penalized Log-Likelihood

Given $n$ samples $\{\boldsymbol{x}(t)\}_{t=0}^{n-1}$ of zero-mean $\boldsymbol{x}$, define the sample covariance $\hat{\boldsymbol{\Sigma}}=\frac{1}{n} \sum_{t=0}^{n-1} \boldsymbol{x}(t) \boldsymbol{x}^{\top}(t)$. Let $\boldsymbol{X}=$ $[\boldsymbol{x}(0) \boldsymbol{x}(1) \cdots \boldsymbol{x}(n-1)]^{\top}$. The log-likelihood, up to some constants, is

$$
\begin{equation*}
\ln f_{\boldsymbol{X}}(\boldsymbol{X})=\ln (|\boldsymbol{\Omega}|)-\operatorname{tr}(\hat{\boldsymbol{\Sigma}} \boldsymbol{\Omega}) \tag{3}
\end{equation*}
$$

To estimate sparse $\boldsymbol{\Omega}$, consider minimization of a penalized version of the negative log-likelihood

$$
\begin{equation*}
L(\boldsymbol{X} ; \boldsymbol{\Omega})=-\ln f_{\boldsymbol{X}}(\boldsymbol{X})+P(\boldsymbol{\Omega}) \tag{4}
\end{equation*}
$$

using a sparse-group SCAD penalty

$$
\begin{equation*}
P(\boldsymbol{\Omega})=\sum_{i \neq j}^{m p} P_{\alpha \lambda}\left(\Omega_{i j}\right)+\sum_{k \neq \ell}^{p} P_{(1-\alpha) \lambda}\left(\left\|\boldsymbol{\Omega}^{(k \ell)}\right\|_{F}\right) \tag{5}
\end{equation*}
$$

where, for some $a>2$, the $\operatorname{SCAD}$ penalty is defined as $P_{\lambda}(\theta)=\lambda|\theta|$ for $|\theta| \leq \lambda,=\frac{2 a \lambda|\theta|-|\theta|^{2}-\lambda^{2}}{2(a-1)}$ for $\lambda<|\theta|<a \lambda$, and $=\frac{\lambda^{2}(a+1)}{2}$ for $|\theta| \geq a \lambda, \lambda>0$ is a tuning parameter used to control sparsity, and $0 \leq \alpha \leq 1$. To explain $\alpha$, first consider sparse-group lasso penalty [17] where $P_{\lambda}(\theta)$ is replaced with

$$
\begin{equation*}
P_{s g l}(\boldsymbol{\Omega})=\alpha \lambda \sum_{i \neq j}^{m p}\left|\Omega_{i j}\right|+(1-\alpha) \lambda \sum_{k \neq \ell}^{p}\left\|\boldsymbol{\Omega}^{(j k)}\right\|_{F} \tag{6}
\end{equation*}
$$

$0 \leq \alpha \leq 1$ yields a convex combination of lasso and group lasso penalties ( $\alpha=0$ gives the group-lasso fit while $\alpha=1$ yields the lasso fit). In SG-SCAD each of the lasso penalties in $P_{s g l}(\boldsymbol{\Omega})$ is replaced with SCAD penalties, mimicking group SCAD of [15].

The first-order derivative of $P_{\lambda}(\theta)$ w.r.t. $|\theta|$ is $P_{\lambda}^{\prime}(\theta)=\lambda$ for $|\theta| \leq \lambda,=\frac{a \lambda-|\theta|}{a-1}$ for $\lambda<|\theta|<a \lambda$, and $=0$ for $|\theta| \geq a \lambda$. Its second-order derivative is $P_{\lambda}^{\prime \prime}(\theta)=\frac{-1}{a-1}$ for $\lambda<|\theta|<a \lambda$, and $=0$ otherwise. The SCAD penalty was proposed by [18] and exploited for graphical model selection in [16]. As suggested in [16], we take $a=3.7$ in this paper.

Algorithm 1 ADMM Algorithm for Sparse-Group Graphical Lasso (typos in [14] corrected)
Input: Number of samples $n$, number of nodes $p$, number of attributes $m$, data $\{\boldsymbol{x}(t)\}_{t=0}^{n-1}, \boldsymbol{x} \in \mathbb{R}^{m p}$, regularization and penalty parameters $\lambda, \alpha$ and $\rho_{0}$, tolerances $\tau_{a b s}$ and $\tau_{\text {rel }}$, variable penalty factor $\mu$, maximum number of iterations $i_{\max }$ Output: estimated inverse covariance $\hat{\Omega}$ and edge-set $\hat{\mathcal{E}}$

Calculate sample covariance $\hat{\boldsymbol{\Sigma}}=\frac{1}{n} \sum_{t=0}^{n-1} \boldsymbol{x}(t) \boldsymbol{x}^{\top}(t)$ (after centering $\boldsymbol{x}(t)$ ).
Initialize: $\boldsymbol{U}^{(0)}=\boldsymbol{W}^{(0)}=\mathbf{0}, \boldsymbol{\Omega}^{(0)}=(\operatorname{diag}(\hat{\boldsymbol{\Sigma}}))^{-1}$, where $\boldsymbol{U}, \boldsymbol{W} \in \mathbb{R}^{(m p) \times(m p)}, \rho^{(0)}=\rho_{0}$
converged $=$ false, $i=0$
while converged $=$ false and $i \leq i_{\max }$, do
Eigen-decompose $\hat{\boldsymbol{\Sigma}}-\rho^{(i)}\left(\overline{\boldsymbol{W}}^{(i)}-\boldsymbol{U}^{(i)}\right)$ as $\hat{\boldsymbol{\Sigma}}-$ $\rho^{(i)}\left(\boldsymbol{W}^{(i)}-\boldsymbol{U}^{(i)}\right)=\boldsymbol{V} \boldsymbol{D} \boldsymbol{V}^{\top}$ with diagonal matrix $\boldsymbol{D}$ consisting of eigenvalues. Define diagonal matrix $\tilde{D}$ with $\ell$ th diagonal element $\tilde{D}_{\ell \ell}=\left(-D_{\ell \ell}+\right.$ $\left.\sqrt{\boldsymbol{D}_{\ell \ell}^{2}+4 \rho^{(i)}}\right) /\left(2 \rho^{(i)}\right)$. Set $\boldsymbol{\Omega}^{(i+1)}=\boldsymbol{V} \tilde{\boldsymbol{D}} \boldsymbol{V}^{\top}$.
Set $\boldsymbol{A}^{(j k)}=\left(\boldsymbol{\Omega}^{(j k)}\right)^{(i+1)}+\left(\boldsymbol{U}^{(j k)}\right)^{(i)}$. Define soft thresholding scalar operator $S(a, \beta):=(1-\beta /|a|)_{+} a$ where $(a)_{+}:=\max (0, a)$. The diagonal $m \times m$ subblocks of $\boldsymbol{W}$ are updated as

$$
\left[\left(\boldsymbol{W}^{(j j)}\right)^{(i+1)}\right]_{s t}=\left\{\begin{array}{c}
{\left[\boldsymbol{A}^{(j j)}\right]_{s s} \quad \text { if } s=t} \\
S\left(\left[\boldsymbol{A}^{(j j)}\right]_{s t}, \frac{\alpha \lambda}{\rho^{(i)}}\right) \text { if } s \neq t
\end{array}\right.
$$

$j=1,2, \cdots, p, s, t=1,2, \cdots, m$. The off-diagonal $m \times m$ subblocks of $\boldsymbol{W}$ are updated as

$$
\left(\boldsymbol{W}^{(j k)}\right)^{(i+1)}=\boldsymbol{B}\left(1-\frac{(1-\alpha) \lambda}{\rho^{(i)}\|\boldsymbol{B}\|_{F}}\right)_{+}
$$

where $\boldsymbol{B}=\boldsymbol{S}\left(\boldsymbol{A}^{(j k)}, \alpha \lambda / \rho^{(i)}\right), \boldsymbol{S}(\boldsymbol{A}, \alpha)$ denotes elementwise matrix soft thresholding, specified by $[\boldsymbol{S}(\boldsymbol{A}, \alpha)]_{s t}:=S\left([\boldsymbol{A}]_{s t}, \alpha\right)$, and $j \neq k=1,2, \cdots, p$. Dual update $\boldsymbol{U}^{(i+1)}=\boldsymbol{U}^{(i)}+\left(\boldsymbol{\Omega}^{(i+1)}-\boldsymbol{W}^{(i+1)}\right)$. Check convergence. Set tolerances

$$
\begin{aligned}
\tau_{\text {pri }} & =m p \tau_{a b s}+\tau_{\text {rel }} \max \left(\left\|\boldsymbol{\Omega}^{(i+1)}\right\|_{F},\left\|\boldsymbol{W}^{(i+1)}\right\|_{F}\right) \\
\tau_{\text {dual }} & =m p \tau_{a b s}+\tau_{\text {rel }}\left\|\boldsymbol{U}^{(i+1)}\right\|_{F} / \rho^{(i)}
\end{aligned}
$$

Define $d_{p}=\left\|\boldsymbol{\Omega}^{(i+1)}-\boldsymbol{W}^{(i+1)}\right\|_{F}$ and $d_{d}=$ $\rho^{(i)}\left\|\boldsymbol{W}^{(i+1)}-\boldsymbol{W}^{(i)}\right\|_{F}$. If $\left(d_{p} \leq \tau_{p r i}\right)$ and $\left(d_{d} \leq\right.$ $\left.\tau_{\text {dual }}\right)$, set converged $=$ true.
9: Update penalty parameter $\rho$ :

$$
\rho^{(i+1)}= \begin{cases}2 \rho^{(i)} & \text { if } d_{p}>\mu d_{d} \\ \rho^{(i)} / 2 & \text { if } d_{d}>\mu d_{p} \\ \rho^{(i)} & \text { otherwise }\end{cases}
$$

We also need to set $\boldsymbol{U}^{(i+1)}=\boldsymbol{U}^{(i+1)} / 2$ for $d_{p}>\mu d_{d}$ and $\boldsymbol{U}^{(i+1)}=2 \boldsymbol{U}^{(i+1)}$ for $d_{d}>\mu d_{p}$.
$i \leftarrow i+1$
end while
For $j \neq k$, if $\left\|\boldsymbol{W}^{(j k)}\right\|_{F}>0$, assign edge $\{j, k\} \in \hat{\mathcal{E}}$, else $\{j, k\} \notin \hat{\mathcal{E}}$. Inverse covariance estimate $\hat{\boldsymbol{\Omega}}=\boldsymbol{W}$.

## III. Solution

Similar to the single attribute results of [16], since SCAD penalty is non-convex, we first solve the SGL problem using [14], and then linearize the SG-SCAD function around the SGL estimate, which then results in a convex problem. We first recall the ADMM-based SGL solution of [14]. Using variable splitting, consider
$\min _{\boldsymbol{\Omega} \succ \mathbf{0}, \boldsymbol{W}}\left\{\operatorname{tr}(\hat{\boldsymbol{\Sigma}} \boldsymbol{\Omega})-\ln (|\boldsymbol{\Omega}|)+P_{\text {sgl }}(\boldsymbol{W})\right\}$ subject to $\boldsymbol{\Omega}=\boldsymbol{W}$.
The scaled augmented Lagrangian for this problem is [21]

$$
\begin{equation*}
L_{\rho}=\operatorname{tr}(\hat{\boldsymbol{\Sigma}} \boldsymbol{\Omega})-\ln (|\boldsymbol{\Omega}|)+P_{\text {sgl }}(\boldsymbol{W})+\frac{\rho}{2}\|\boldsymbol{\Omega}-\boldsymbol{W}+\boldsymbol{U}\|_{F}^{2} \tag{7}
\end{equation*}
$$

where $\boldsymbol{U}$ is the dual variable, and $\rho>0$ is a penalty parameter. The ADMM-based solution of [14] is given in Algorithm 1 (with typos in [14] corrected), where we use the convergence criterion following [21, Sec. 3.3.1] and varying penalty parameter $\rho$ following [21, Sec. 3.4.1]. At $(i+1)$ st iteration, the primal residual is given by $\boldsymbol{\Omega}^{(i+1)}-\boldsymbol{W}^{(i+1)}$ and the dual residual by $\rho^{(i)}\left(\boldsymbol{W}^{(i+1)}-\boldsymbol{W}^{(i)}\right)$. Convergence criterion is met when the norms of these residuals are below tolerances $\tau_{p r i}$ and $\tau_{\text {dual }}$, respectively; see line 8 of Algorithm 1. In turn, $\tau_{p r i}$ and $\tau_{\text {dual }}$ are chosen using an absolute and relative criterion as in line 8 of Algorithm 1 where $\tau_{a b s}$ and $\tau_{r e l}$ are user chosen absolute and relative tolerances, respectively.

Use Algorithm 1 to obtain SGL solution $\hat{\boldsymbol{\Omega}}^{(1)}$, $\hat{\boldsymbol{W}}^{(1)}$ and $\hat{\boldsymbol{U}}^{(1)}$ to (7). Linearize $P(\boldsymbol{\Omega})$ around $\hat{\boldsymbol{W}}^{(1)}$ as

$$
\begin{align*}
P_{l i n}(\boldsymbol{W}) & =\sum_{i \neq j}^{m p} P_{\alpha \lambda}^{\prime}\left(\hat{W}_{i j}^{(1)}\right)\left|W_{i j}\right| \\
& +\sum_{k \neq \ell}^{p} P_{(1-\alpha) \lambda}^{\prime}\left(\left\|\left(\hat{\boldsymbol{W}}^{(1)}\right)^{(k \ell)}\right\|_{F}\right)\left\|\boldsymbol{W}^{(j k)}\right\|_{F} \tag{8}
\end{align*}
$$

Again solve a convex SGL problem after replacing $P_{s g l}(\boldsymbol{W})$ with $P_{\text {lin }}(\boldsymbol{W})$, and with following "obvious" modifications to Algorithm 1: in line 6 therein, replace $\alpha \lambda$ with $P_{\alpha \lambda}^{\prime}\left(\hat{W}_{i j}^{(1)}\right)$, and replace $(1-\alpha) \lambda$ with $P_{(1-\alpha) \lambda}^{\prime}\left(\left\|\left(\hat{\boldsymbol{W}}^{(1)}\right)^{(k \ell)}\right\|_{F}\right)$. Recall that $P_{\lambda}^{\prime}(\theta)=\lambda$ for $|\theta| \leq \lambda,=\frac{a \lambda-|\theta|}{a-1}$ for $\lambda<|\theta|<a \lambda$, and $=0$ for $|\theta| \geq a \lambda$. The resulting (SG-SCAD) solution is denoted by $\hat{\boldsymbol{\Omega}}^{(2)}, \hat{\boldsymbol{W}}^{(2)}$ and $\hat{\boldsymbol{U}}^{(2)}$.

## IV. Theoretical Analysis

Let $\Omega_{0}$ denote the true $\Omega$ and $\mathcal{E}_{0}$ denote the true edgeset. Assume
(A1) $\operatorname{Card}\left(\mathcal{E}_{0}\right)=\left|\mathcal{E}_{0}\right| \leq s_{n 0}$.
(A2) $0<\beta_{\min } \leq \phi_{\min }\left(\boldsymbol{\Sigma}_{\mathbf{0}}\right) \leq \phi_{\max }\left(\boldsymbol{\Sigma}_{\mathbf{0}}\right) \leq \beta_{\max }<\infty$ where $\boldsymbol{\Sigma}_{0}=\boldsymbol{\Omega}_{0}^{-1}$, and $\beta_{\min }$ and $\beta_{\max }$ are not functions of $n$.
(A3) $\min _{\left\{\{i, j\}: \Omega_{0 i j} \neq 0\right\}}\left|\Omega_{0 i j}\right| \geq \delta_{0}>0$.
Let $\hat{\boldsymbol{\Omega}}_{\lambda}=\arg \min _{\boldsymbol{\Omega} \succ \mathbf{0}} L(\boldsymbol{X} ; \boldsymbol{\Omega})$. We denote $p$ by $p_{n}$ to indicate that it can grow with $n$.
Theorem 1 (Consistency): For $\tau>2$, let

$$
\begin{equation*}
C_{0}=40 \max _{k}\left(\Sigma_{0 k k}\right) \sqrt{2\left(\tau+\ln (4) / \ln \left(m p_{n}\right)\right)} \tag{9}
\end{equation*}
$$

For $\delta_{1} \in(0,1)$ and "small" $\delta_{2}>0$, let

$$
\begin{align*}
& M=\left(1+\delta_{1}\right)^{2}\left(2+\delta_{2}\right) C_{0} / \beta_{\min }^{2}  \tag{10}\\
& r_{n}=\sqrt{\frac{\left(m p_{n}+m^{2} s_{n 0}\right) \ln \left(m p_{n}\right)}{n}}=o(1)  \tag{11}\\
& N_{1}=2\left(\ln (4)+\tau \ln \left(m p_{n}\right)\right),  \tag{12}\\
& N_{2}=\arg \min \left\{n: r_{n} \leq \frac{\delta_{1} \beta_{\min }}{\left(1+\delta_{1}\right)^{2}\left(2+\delta_{2}\right) C_{0}}\right\} \tag{13}
\end{align*}
$$

Pick $\lambda_{n}$ and integer $N_{3}$ as ( $a>2$ is a SCAD parameter)

$$
\begin{align*}
& \lambda_{n}= \begin{cases}\max \left(\frac{1}{\alpha}, \frac{1}{1-\alpha}\right) \max \left(M, C_{0}\right) r_{n}, & \alpha \in(0,1) \\
\max \left(M, C_{0}\right) r_{n}, & \alpha=0 \text { or } 1\end{cases}  \tag{14}\\
& N_{3}=\arg \min \left\{n: \lambda_{n}<\frac{\min _{\left\{\{i, j\}: \Omega_{0 i j} \neq 0\right\}}\left|\Omega_{0 i j}\right|}{a}\right\} . \tag{15}
\end{align*}
$$

For $n>\max \left\{N_{1}, N_{2}, N_{3}\right\}$, under assumptions (A1)-(A3), there exists a local minimizer $\hat{\boldsymbol{\Omega}}_{\lambda}$ such that

$$
\begin{equation*}
\left\|\hat{\boldsymbol{\Omega}}_{\lambda}-\boldsymbol{\Omega}_{0}\right\|_{F} \leq M r_{n} \tag{16}
\end{equation*}
$$

with probability $>1-1 /\left(m p_{n}\right)^{\tau-2}$. In terms of rate of convergence, $\left\|\hat{\boldsymbol{\Omega}}_{\lambda}-\boldsymbol{\Omega}_{0}\right\|_{F}=\mathcal{O}_{P}\left(r_{n}\right)$

## V. Proof of Theorem 1

Lemma 1 follows from [20, Lemma 1].
Lemma 1: Under Assumption (A2), the sample covariance $\hat{\boldsymbol{\Sigma}}$ satisfies the tail bound

$$
\begin{equation*}
P\left(\max _{k, \ell}\left|\left[\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right]_{k l}\right|>C_{0} \sqrt{\frac{\ln \left(m p_{n}\right)}{n}}\right) \leq \frac{1}{\left(m p_{n}\right)^{\tau-2}} \tag{17}
\end{equation*}
$$

for $\tau>2$, if the sample size $n>N_{1}$, where $C_{0}$ is defined in (9) and $N_{1}$ is defined in (12).

We now turn to the proof of Theorem 1.
Proof of Theorem 1. Let $\boldsymbol{\Omega}=\boldsymbol{\Omega}_{0}+\boldsymbol{\Delta}$ with both $\boldsymbol{\Omega}, \boldsymbol{\Omega}_{0} \succ \mathbf{0}$, and $Q(\boldsymbol{\Omega}):=L(\boldsymbol{X} ; \boldsymbol{\Omega})-L\left(\boldsymbol{X} ; \boldsymbol{\Omega}_{0}\right)$. The estimate $\hat{\boldsymbol{\Omega}}_{\lambda}$, denoted by $\hat{\boldsymbol{\Omega}}$ hereafter suppressing dependence upon $\lambda$, minimizes $Q(\boldsymbol{\Omega})$, or equivalently, $\hat{\boldsymbol{\Delta}}=\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_{0}$ minimizes $G(\boldsymbol{\Delta}):=Q\left(\boldsymbol{\Omega}_{0}+\boldsymbol{\Delta}\right)$. We will follow, for the most part, the method of proof of [19, Theorem 1] pertaining to lasso penalty. Consider the set

$$
\begin{equation*}
\Theta_{n}(M):=\left\{\boldsymbol{\Delta}: \boldsymbol{\Delta}=\boldsymbol{\Delta}^{\top},\|\boldsymbol{\Delta}\|_{F}=M r_{n}\right\} \tag{18}
\end{equation*}
$$

where $M$ and $r_{n}$ are as in (10) and (11), respectively. Since $G(\hat{\boldsymbol{\Delta}}) \leq G(\mathbf{0})=0$, if we can show that $\inf _{\boldsymbol{\Delta}}\{G(\boldsymbol{\Delta}):$ $\left.\boldsymbol{\Delta} \in \Theta_{n}(M)\right\}>0$, then the minimizer $\hat{\boldsymbol{\Delta}}$ must be inside $\Theta_{n}(M)$, and hence $\|\hat{\boldsymbol{\Delta}}\|_{F} \leq M r_{n}$. It is shown in [19, (9)] that $\ln \left(\left|\boldsymbol{\Omega}_{0}+\boldsymbol{\Delta}\right|\right)-\ln \left(\left|\boldsymbol{\Omega}_{0}\right|\right)=\operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathbf{0}} \boldsymbol{\Delta}\right)-A_{1}$ where, with $\boldsymbol{H}\left(\boldsymbol{\Omega}_{0}, \boldsymbol{\Delta}, v\right)=\left(\boldsymbol{\Omega}_{\mathbf{0}}+v \boldsymbol{\Delta}\right)^{-1} \otimes\left(\boldsymbol{\Omega}_{\mathbf{0}}+v \boldsymbol{\Delta}\right)^{-1}$ and $v$ denoting a scalar,

$$
\begin{equation*}
A_{1}:=\operatorname{vec}(\boldsymbol{\Delta})^{\top}\left(\int_{0}^{1}(1-v) \boldsymbol{H}\left(\boldsymbol{\Omega}_{0}, \boldsymbol{\Delta}, v\right) d v\right) \operatorname{vec}(\boldsymbol{\Delta}) \tag{19}
\end{equation*}
$$

Noting that $\boldsymbol{\Omega}^{-1}=\boldsymbol{\Sigma}$ and setting $\bar{\lambda}_{1}=\alpha \lambda_{n}$ and $\bar{\lambda}_{2}=(1-$ $\alpha) \lambda_{n}$, we can rewrite $G(\boldsymbol{\Delta})$ as

$$
\begin{align*}
& G(\boldsymbol{\Delta})=\sum_{i=1}^{4} A_{i}, \quad A_{2}:=\operatorname{tr}\left(\left(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right) \boldsymbol{\Delta}\right)  \tag{20}\\
& A_{3}:=\sum_{i \neq j}^{m p_{n}}\left(P_{\bar{\lambda}_{1}}\left(\Omega_{0 i j}+\Delta_{i j}\right)-P_{\bar{\lambda}_{1}}\left(\Omega_{0 i j}\right)\right)  \tag{21}\\
& A_{4}:=\sum_{k \neq \ell}^{p_{n}}\left(P_{\bar{\lambda}_{2}}\left(\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}+\boldsymbol{\Delta}^{(k \ell)}\right\|_{F}\right)-P_{\bar{\lambda}_{2}}\left(\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}\right\|_{F}\right)\right) \tag{22}
\end{align*}
$$

Following [19, p. 502], we have

$$
\begin{equation*}
A_{1} \geq \frac{\|\boldsymbol{\Delta}\|_{F}^{2}}{2\left(\left\|\boldsymbol{\Omega}_{0}\right\|+\|\boldsymbol{\Delta}\|\right)^{2}} \geq \frac{\|\boldsymbol{\Delta}\|_{F}^{2}}{2\left(\beta_{\min }^{-1}+M r_{n}\right)^{2}} \tag{23}
\end{equation*}
$$

where we have used the fact that $\left\|\boldsymbol{\Omega}_{0}\right\|=\left\|\boldsymbol{\Sigma}_{0}^{-1}\right\|=$ $\phi_{\max }\left(\boldsymbol{\Sigma}_{0}^{-1}\right)=\left(\phi_{\min }\left(\boldsymbol{\Sigma}_{0}\right)\right)^{-1} \leq \beta_{\min }^{-1}$ and $\|\boldsymbol{\Delta}\| \leq\|\boldsymbol{\Delta}\|_{F}=$ $M r_{n}=\mathcal{O}\left(r_{n}\right)$. We now consider $A_{2}$ in (20). We have

$$
\begin{align*}
A_{2} & =L_{21}+L_{22}, \quad L_{22}=\sum_{\{i, j\} \in \overline{\mathcal{E}}_{0}^{c}}\left[\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right]_{i j} \Delta_{j i}  \tag{24}\\
L_{21} & =\sum_{\{i, j\} \in \overline{\mathcal{E}}_{0}}\left[\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right]_{i j} \Delta_{j i}+\sum_{i}\left[\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right]_{i i} \Delta_{i i} \tag{25}
\end{align*}
$$

To bound $L_{21}$, using Cauchy-Schwartz inequality and Lemma 1 , with probability $>1-1 /\left(m p_{n}\right)^{\tau-2}$,

$$
\begin{align*}
& \left|L_{21}\right| \leq\left\|\boldsymbol{\Delta}_{\overline{\mathcal{E}}_{0}}^{-}+\boldsymbol{\Delta}^{+}\right\|_{1} \max _{i, j}\left|\left[\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right]_{i j}\right| \\
& \quad \leq \sqrt{m^{2} s_{n 0}+m p_{n}}\|\boldsymbol{\Delta}\|_{F} C_{0} \sqrt{\ln \left(m p_{n}\right) / n}=C_{0}\|\boldsymbol{\Delta}\|_{F} r_{n} \tag{26}
\end{align*}
$$

We consider $L_{22}$ later as a part of $A_{3}$ where

$$
\begin{align*}
A_{3} & =L_{31}+L_{32}, \quad L_{32}=\sum_{\{i, j\} \in \overline{\mathcal{E}}_{0}^{c}} P_{\bar{\lambda}_{1}}\left(\Delta_{i j}\right)  \tag{27}\\
L_{31} & =\sum_{\{i, j\} \in \overline{\mathcal{E}}_{0}}\left(P_{\bar{\lambda}_{1}}\left(\Omega_{0 i j}+\Delta_{i j}\right)-P_{\bar{\lambda}_{1}}\left(\Omega_{0 i j}\right)\right) . \tag{28}
\end{align*}
$$

For $\lambda_{n}$ as in (14), $\bar{\lambda}_{1} \geq M r_{n} \geq\left|\Delta_{i j}\right|\left(\right.$ since $\left.\|\boldsymbol{\Delta}\|_{F}=M r_{n}\right)$, leading to $P_{\bar{\lambda}_{1}}\left(\Delta_{i j}\right)=\alpha \lambda_{n}\left|\Delta_{i j}\right|$. Consider $L_{32}$ with $\alpha L_{22}$

$$
\begin{align*}
L_{32}-\alpha\left|L_{22}\right| & \geq \sum_{\{i, j\} \in \overline{\mathcal{E}}_{0}^{c}}\left(\alpha \lambda_{n}\left|\Delta_{i j}\right|-\left|\left[\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right]_{i j}\right|\left|\Delta_{i j}\right|\right) \\
& \geq \alpha \lambda_{n}\left(1-\frac{C_{0}}{\alpha \lambda_{n}} \sqrt{\frac{\ln \left(m p_{n}\right)}{n}}\right) \sum_{\{i, j\} \in \overline{\mathcal{E}}_{0}^{c}}\left|\Delta_{i j}\right|>0 \tag{29}
\end{align*}
$$

with prob. $>1-1 /\left(m p_{n}\right)^{\tau-2}$, since $\frac{C_{0}}{\alpha \lambda_{n}} \sqrt{\ln \left(m p_{n}\right) / n}<1$. Now we bound $\left|L_{31}\right|$. A Taylor series expansion of $P_{\lambda}(\theta)$ for $\theta>0$, around $\theta_{0}>0$, is given by $P_{\lambda}(\theta)=P_{\lambda}\left(\theta_{0}\right)+$ $P_{\lambda}^{\prime}\left(\theta_{0}\right)\left(\theta-\theta_{0}\right)+P_{\lambda}^{\prime \prime}(\tilde{\theta}) \frac{\left(\theta-\theta_{0}\right)^{2}}{2}$ where $\tilde{\theta}=\theta_{0}+\gamma\left(\theta-\theta_{0}\right)$ for some $\gamma \in[0,1]$. Setting $\lambda=\bar{\lambda}_{1}, \theta_{0}=\left|\Omega_{0 i j}\right|$ and $\theta=$ $\left|\Omega_{0 i j}+\Delta_{i j}\right|$, and noting that $P_{\lambda}^{\prime \prime}(\tilde{\theta}) \leq 0$ for any $\tilde{\theta}>0$, and $\left|\Omega_{0 i j}\right|>0$ for $\{i, j\} \in \overline{\mathcal{E}}_{0}$, we have $P_{\bar{\lambda}_{1}}\left(\Omega_{0 i j}+\Delta_{i j}\right) \leq$
$P_{\bar{\lambda}_{1}}\left(\Omega_{0 i j}\right)+P_{\bar{\lambda}_{1}}^{\prime}\left(\Omega_{0 i j}\right)\left(\left|\Omega_{0 i j}+\Delta_{i j}\right|-\left|\Omega_{0 i j}\right|\right)$. Since $P_{\bar{\lambda}_{1}}^{\prime}(\theta) \geq$ $0 \forall \theta, P_{\bar{\lambda}_{1}}^{\prime}\left(\left|\Omega_{0 i j}\right|\right)=0$ for $n \geq N_{3},\{i, j\} \in \overline{\mathcal{E}}_{0}$,

$$
\begin{align*}
\left|L_{31}\right| & \leq \sum_{\{i, j\} \in \overline{\mathcal{E}}_{0}} P_{\bar{\lambda}_{1}}^{\prime}\left(\left|\Omega_{0 i j}\right|\right)| | \Omega_{0 i j}+\Delta_{i j}\left|-\left|\Omega_{0 i j}\right|\right| \\
& =0 \text { for } n \geq N_{3} \tag{30}
\end{align*}
$$

Now consider $A_{4}$ which can be expressed as

$$
\begin{aligned}
& A_{4}=L_{41}+L_{42}, \quad L_{42}=\sum_{\{k, \ell\} \in \mathcal{E}_{0}^{c}} P_{\bar{\lambda}_{2}}\left(\boldsymbol{\Delta}^{(k \ell)} \|_{F}\right) \\
& L_{41}=\sum_{\{k, \ell\} \in \mathcal{E}_{0}}\left(P_{\bar{\lambda}_{2}}\left(\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}+\boldsymbol{\Delta}^{(k \ell)}\right\|_{F}\right)-P_{\bar{\lambda}_{1}}\left(\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}\right\|_{F}\right)\right) .
\end{aligned}
$$

Similar to $\left|L_{31}\right|$, we have

$$
\begin{align*}
\left|L_{41}\right| \leq & \sum_{\{k, \ell\} \in \mathcal{E}_{0}} P_{\bar{\lambda}_{2}}^{\prime}\left(\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}\right\|_{F}\right) \mid\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}+\boldsymbol{\Delta}^{(k \ell)}\right\|_{F} \\
& -\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}\right\|_{F} \mid=0 \text { for } n \geq N_{3} \tag{32}
\end{align*}
$$

since $P_{\bar{\lambda}_{2}}^{\prime}\left(\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}\right\|_{F}\right)=0$ for $n \geq N_{3}$ if $\{k, \ell\} \in \mathcal{E}_{0}$ and since $\min _{k, \ell}\left\|\boldsymbol{\Omega}_{0}^{(k \ell)}\right\|_{F} \geq \min _{i, j}\left|\Omega_{0 i j}\right|$. Now consider $L_{42}$ with (1$\alpha) L_{22}$. With $u=(k-1) m+s$ and $v=(l-1) m+t$, we have

$$
\begin{aligned}
& L_{42}-(1-\alpha)\left|L_{22}\right| \geq \sum_{\{k, \ell\} \in \mathcal{E}_{0}^{c}}\left((1-\alpha) \lambda_{n}\left\|\boldsymbol{\Delta}^{(k \ell)}\right\|_{F}\right. \\
& \left.\quad-(1-\alpha) \sum_{s, t=1}^{m}\left|\left[\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right]_{u v}\right|\left|\Delta_{u v}\right|\right) \\
& \geq(1-\alpha) \sum_{\{k, \ell\} \in \mathcal{E}_{0}^{c}}\left(\lambda_{n}\left\|\boldsymbol{\Delta}^{(k \ell)}\right\|_{F}-m C_{0} \sqrt{\frac{\ln \left(m p_{n}\right)}{n}}\left\|\boldsymbol{\Delta}^{(k \ell)}\right\|_{F}\right) \\
& \geq(1-\alpha) \lambda_{n}\left(1-\frac{m C_{0}}{(1-\alpha) \lambda_{n}} \sqrt{\frac{\ln \left(m p_{n}\right)}{n}}\right) \sum_{\{k, \ell\} \in \mathcal{E}_{0}^{c}}\left\|\boldsymbol{\Delta}^{(k \ell)}\right\|_{F}
\end{aligned}
$$

$$
\begin{equation*}
>0 \tag{33}
\end{equation*}
$$

with prob. $>1-1 /\left(m p_{n}\right)^{\tau-2}$, since $\frac{m C_{0}}{(1-\alpha) \lambda_{n}} \sqrt{\ln \left(m p_{n}\right) / n}<$ 1. Combining $A_{2}, A_{3}$ and $A_{4}$, we have

$$
\begin{align*}
& A_{2}+A_{3}+A_{4}=\sum_{i=1}^{3} \sum_{j=1}^{2} L_{i j} \geq-\left|L_{21}\right|+L_{32}-\alpha\left|L_{22}\right| \\
& \quad+L_{31}+L_{42}-(1-\alpha)\left|L_{22}\right|+L_{41} \\
& \geq-\left|L_{21}\right|+L_{31}+L_{41} \geq C_{0}\|\boldsymbol{\Delta}\|_{F} r_{n} \text { for } n \geq N_{3} \tag{34}
\end{align*}
$$

where we have used (26), (29), (30), (32) and (33). Using (20), the bound (23) on $A_{1}$ and (34) on $A_{2}+A_{3}+A_{4}$, and $\|\boldsymbol{\Delta}\|_{F}=M r_{n}$, we have with probability $>1-1 /\left(m p_{n}\right)^{\tau-2}$,

$$
\begin{equation*}
G(\boldsymbol{\Delta}) \geq\|\boldsymbol{\Delta}\|_{F}^{2}\left[\frac{1}{2\left(\beta_{\min }^{-1}+M r_{n}\right)^{2}}-\frac{C_{0}}{M}\right] \tag{35}
\end{equation*}
$$

For $n \geq N_{2}$, if we pick $M$ as specified in (10), we obtain $M r_{n} \leq M r_{N_{2}} \leq \delta_{1} / \beta_{\text {min }}$. Then

$$
\frac{1}{2\left(\beta_{\min }^{-1}+M r_{n}\right)^{2}} \geq \frac{\beta_{\min }^{2}}{2\left(1+\delta_{1}\right)^{2}}=\frac{\left(2+\delta_{2}\right) C_{0}}{2 M}>\frac{C_{0}}{M}
$$

implying $G(\boldsymbol{\Delta})>0$. For $\alpha=0$, omit $A_{3}$, and for $\alpha=1$, omit $A_{4}$ from $G(\boldsymbol{\Delta})$, to get $G(\boldsymbol{\Delta})>0$, completing the proof.

## VI. Simulation Example

Now we consider an Erdös-Rényi graph where $p$ nodes are connected to each other with probability $p_{e r}=0.05$. In the upper triangular $\bar{\Omega}$, using the notation of (2), we set $\left[\overline{\mathbf{\Omega}}^{(j k)}\right]_{s t}=0.5^{|s-t|}$ for $j=k=1, \cdots, p, s, t=1, \cdots, m$. For $j \neq k$, if the two nodes are not connected, we have $\overline{\boldsymbol{\Omega}}^{(j k)}=\mathbf{0}$, and if nodes $j$ and $k$ are connected in the chain graph, then $\left[\bar{\Omega}^{(j k)}\right]_{s t}$ is uniformly distributed over $[-0.4,-0.1] \cup[0.1,0.4]$ if $s \neq t$, and $\left[\bar{\Omega}^{(j k)}\right]_{s t}=0$ if $s=t$. Now add $\gamma \boldsymbol{I}$ to $\boldsymbol{\Omega}$ with $\gamma$ picked to make minimum eigenvalue of $\boldsymbol{\Omega}+\gamma \boldsymbol{I}$ equal to 0.5 . With $\boldsymbol{\Phi} \boldsymbol{\Phi}^{\top}=(\boldsymbol{\Omega}+\gamma \boldsymbol{I})^{-1}$, we generate $\boldsymbol{x}=\boldsymbol{\Phi} \boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{R}^{m p}$ as Gaussian $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. We generate $n$ i.i.d. observations from $\boldsymbol{x}$, with $m=3$, $p=400, n \in\{100,200,400,800,1600,3200\}$. We then have $\frac{1}{2} \mathbb{E}\{|\mathcal{E}|\}=3990$.


Fig. 1: Error norm $\left\|\hat{\boldsymbol{W}}-\boldsymbol{\Omega}_{0}\right\|_{F} /\left\|\boldsymbol{\Omega}_{0}\right\|_{F}$ and corresponding $F_{1}$ values; $p=400$.

We used the solution outlined in Sec. III with $a=3.7$, $\rho_{0}=2, \mu=10$, and $\tau_{a b s}=\tau_{\text {rel }}=10^{-4}$. Simulation results based on 100 runs are shown in Fig. 1 for $p=400$, with varying $n$. We compare the SG-SCAD solution with the SGL solution. The performance metrics used are the $F_{1}$-score and the Frobenius error norm $=\left\|\hat{\boldsymbol{W}}-\boldsymbol{\Omega}_{0}\right\|_{F} /\left\|\boldsymbol{\Omega}_{0}\right\|_{F}$ where $\hat{\boldsymbol{W}}=\hat{\boldsymbol{W}}^{(2)}$ for SG-SCAD penalty and $\hat{\boldsymbol{W}}=\hat{\boldsymbol{W}}^{(1)}$ for SGL penalty. We first selected the tuning parameters $(\lambda, \alpha)$ by searching over a two-dimensional grid to minimize the Hamming distance between $\mathcal{E}_{0}$ and $\hat{\mathcal{E}}$, for $(p, n)=(400,400)$, resulting in $(\lambda, \alpha)=(0.08,0.05)$ for both methods. (In practice, one would use an information criterion.) Then for other values of $n$ (and $p, m$ ), we scale $\lambda$ as $\lambda_{n} \propto \sqrt{s_{n 0}+\left(p_{n} / m\right)}$. $m \cdot \sqrt{\ln \left(m p_{n}\right) / n}$ for SG-SCAD based on (11) and (14), and as $\lambda_{n} \propto m \cdot \sqrt{\ln \left(m p_{n}\right) / n}$ for SGL [14]. It is seen from Fig. 1 that while $F_{1}$ values are comparable, the SG-SCAD approach yields significantly smaller errors in estimating $\boldsymbol{\Omega}$ compared to the SGL approach.

## VII. CONCLUSIONS

We considered the problem of inferring the conditional independence graph of high-dimensional Gaussian vectors from multi-attribute data. We analyzed an SG-SCAD-penalized loglikelihood based objective function to establish consistency of a local estimator of the inverse covariance in a neighborhood of
the true value. An ADMM algorithm based iterative reweighting method was used to optimize the objective function, starting with the globally convergent SGL method of [14]. A numerical example was presented to illustrate the advantage of SG-SCAD over the "usual" SGL penalty.

## REFERENCES

[1] S.L. Lauritzen, Graphical models. Oxford, UK: Oxford Univ. Press, 1996.
[2] J. Whittaker, Graphical Models in Applied Multivariate Statistics. New York: Wiley, 1990.
[3] P. Danaher, P. Wang and D.M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," J. Royal Statistical Society, Series B, vol. 76, pp. 373-397, 2014.
[4] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," Ann. Statist., vol. 34, no. 3, pp. 1436-1462, 2006.
[5] K. Mohan, P. London, M. Fazel, D. Witten and S.I. Lee, "Node-based learning of multiple Gaussian graphical models," J. Machine Learning Research, vol. 15, 2014.
[6] J. Friedman, T. Hastie and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," Biostatistics, vol. 9, no. 3, pp. 432-441, July 2008.
[7] O. Banerjee, L.E. Ghaoui and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," J. Machine Learning Research, vol. 9, pp. 485-516, 2008.
[8] J. Chiquet, G. Rigaill and M. Sundquist, "A multiattribute Gaussian graphical model for inferring multiscale regulatory networks: an application in breast cancer," in Gene Regulatory Networks. Methods in Molecular Biology, G. Sanguinetti and V. Huynh-Thu, Eds., vol 1883. Humana Press, New York, NY, 2019, pp. 143-160.
[9] M. Kolar, H. Liu and E.P. Xing, "Markov network estimation from multi-attribute data," in Proc. 30th Intern. Conf. Machine Learning (ICML), Atlanta, GA, 2013.
[10] M. Kolar, H. Liu and E.P. Xing, "Graph estimation from multi-attribute data," J. Machine Learning Research, vol. 15, pp. 1713-1750, 2014.
[11] J.K. Tugnait, "Deviance tests for graph estimation from multi-attribute Gaussian data," IEEE Trans. Signal Process., vol. 68, pp. 5632-5647, 2020.
[12] E. Pavez and A. Ortega, "Generalized Laplacian precision matrix estimation for graph signal processing," in Proc. IEEE ICASSP 2016, Shanghai, China, March 2016, pp. 6350-6354.
[13] E. Pavez, H.E. Egilmez and A. Ortega, "Learning graphs with monotone topology properties and multiple connected components," IEEE Trans. Signal Process., vol. 66, no. 9, pp. 2399-2413, May 1, 2018.
[14] J.K. Tugnait, "Sparse-group lasso for graph learning from multiattribute data," IEEE Trans. Signal Process., vol. 69, pp. 1771-1786, 2021.
[15] P. Breheny and J. Huang, "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors," Statistics and Computing, vol. 25, pp. 173-187, 2015.
[16] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," Ann. Statist., vol. 37, no. 6B, pp. 42544278, 2009.
[17] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, "A sparse-group lasso," J. Computational Graphical Statistics, vol. 22, pp. 231-245, 2013.
[18] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," J. American Statistical Assoc., vol. 96, pp. 1348-1360, Dec. 2001.
[19] A.J. Rothman, P.J. Bickel, E. Levina and J. Zhu, "Sparse permutation invariant covariance estimation," Electronic J. Statistics, vol. 2, pp 494-515, 2008.
[20] P. Ravikumar, M.J. Wainwright, G. Raskutti and B. Yu, "Highdimensional covariance estimation by minimizing $\ell_{1}$-penalized logdeterminant divergence," Electronic J. Statistics, vol. 5, pp. 935-980, 2011.
[21] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends in Machine Learning, vol. 3, no. 1, pp. 1-122, 2010.


[^0]:    This work was supported by NSF Grants CCF-1617610 and ECCS2040536. Author's email: tugnajk @ auburn.edu

