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

Jitendra K. Tugnait Dept. of Electrical & Computer Eng. 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-penality 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, \dots, x_p , $(\boldsymbol{x} = [x_1 \ x_2 \ \dots \ x_p]^\top)$, is represented using an undirected graph $\mathcal{G} = (V, \mathcal{E})$, where $V = \{1, 2, \dots, p\} =$ [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 x is multivariate Gaussian. Suppose x has positive-definite covariance matrix Σ with inverse covariance matrix $\Omega = \Sigma^{-1}$. Then Ω_{ij} , the (i, j)-th element of Ω , is zero iff x_i and x_j are conditionally independent. Given n samples of x, in highdimensional settings, one estimates Ω 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

This work was supported by NSF Grants CCF-1617610 and ECCS-2040536. Author's email: tugnajk@auburn.edu

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 $A \in \mathbb{R}^{p \times p}$, we use $\phi_{\min}(A)$, $\phi_{\max}(A)$, $|\mathbf{A}|$ and tr(A) to denote the minimum eigenvalue, maximum eigenvalue, determinant and trace of A, respectively. For $B \in \mathbb{R}^{p \times q}$, we have $\|B\| = \sqrt{\phi_{\max}(B^\top B)}$, $\|B\|_F = \sqrt{\operatorname{tr}(B^\top B)}$ and $\|B\|_1 = \sum_{i,j} |B_{ij}|$ where B_{ij} is the (i, j)-th element of B (also denoted by $[B]_{ij}$). Given $A \in \mathbb{R}^{p \times p}$, $A^+ = \operatorname{diag}(A)$ is a diagonal matrix with the same diagonal as A, and $A^- = A - A^+$ is A with all its diagonal elements set to zero. The notation $y_n = \mathcal{O}_P(x_n)$ for random vectors $y_n, x_n \in \mathbb{R}^p$ means that for any $\varepsilon > 0$, there exists $0 < M < \infty$ such that $P(\|y_n\| \le M\|x_n\|) \ge 1 - \varepsilon \ \forall n \ge 1$.

II. SYSTEM MODEL

We will call \mathcal{G} considered in Sec. I a single-attribute graphical model for x. Now consider p jointly Gaussian random vectors $z_i \in \mathbb{R}^m$, $i = 1, 2, \dots, p$. We associate 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 $\{z_i, i \in V\}$. As in the scalar case (m = 1), there is no edge between node *i* and node *j* in \mathcal{G} iff random vectors z_i and 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 *mp*-vector

$$\boldsymbol{x} = [\boldsymbol{z}_1^\top \ \boldsymbol{z}_2^\top \ \cdots \ \boldsymbol{z}_p^\top]^\top \in \mathbb{R}^{mp} \,. \tag{1}$$

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 $(\mathbb{E}\{\mathbf{x}\mathbf{x}^{\top}\})^{-1}$ 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 $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$, where $\bar{V} = [1, mp]$ and

 $\bar{\mathcal{E}} \subseteq \bar{V} \times \bar{V}$. Now $[\boldsymbol{z}_j]_{\ell}$, the ℓ 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 $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$. Corresponding to the edge $\{j, k\} \in \mathcal{E}$ in the multi-attribute $\mathcal{G} = (V, \mathcal{E})$, there are m^2 edges $\{q, r\} \in \bar{\mathcal{E}}$ 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 $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$ is a single-attribute graph. In order for $\bar{\mathcal{G}}$ to reflect the conditional independencies encoded in \mathcal{G} , we must have the equivalence $\{j, k\} \notin \mathcal{E} \Leftrightarrow \bar{\mathcal{E}}^{(jk)} \cap \bar{\mathcal{E}} = \emptyset$, where $\bar{\mathcal{E}}^{(jk)} = \{\{q, r\} \mid q = (j-1)m + s, r = (k-1)m + t, s, t = 1, 2, \cdots, m\}$. Let $\boldsymbol{R}_{xx} = \mathbb{E}\{\boldsymbol{xx}^{\top}\} \succ \mathbf{0}$ and $\boldsymbol{\Omega} = \boldsymbol{R}_{xx}^{-1}$. Define the (j, k)th $m \times m$ subblock $\boldsymbol{\Omega}^{(jk)}$ of $\boldsymbol{\Omega}$ as

$$\mathbf{\Omega}^{(jk)}]_{st} = [\mathbf{\Omega}]_{(j-1)m+s,(k-1)m+t}, \ s, t = 1, 2, \cdots, m.$$
(2)

It is established in [10, Sec. 2.1] that $\Omega^{(jk)} = \mathbf{0} \Leftrightarrow \{j, k\} \notin \mathcal{E}$. Since $\Omega^{(jk)} = \mathbf{0}$ is equivalent to $[\Omega]_{qr} = 0$ for every $\{q, r\} \in \overline{\mathcal{E}}^{(jk)}$, and since, by [1, Proposition 5.2], $[\Omega]_{qr} = 0$ 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

$$\ln f_{\boldsymbol{X}}(\boldsymbol{X}) = \ln(|\boldsymbol{\Omega}|) - \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}).$$
(3)

To estimate sparse Ω , consider minimization of a penalized version of the negative log-likelihood

$$L(\boldsymbol{X};\boldsymbol{\Omega}) = -\ln f_{\boldsymbol{X}}(\boldsymbol{X}) + P(\boldsymbol{\Omega})$$
(4)

using a sparse-group SCAD penalty

$$P(\mathbf{\Omega}) = \sum_{i \neq j}^{mp} P_{\alpha\lambda}(\Omega_{ij}) + \sum_{k \neq \ell}^{p} P_{(1-\alpha)\lambda}(\|\mathbf{\Omega}^{(k\ell)}\|_F), \quad (5)$$

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

$$P_{sgl}(\mathbf{\Omega}) = \alpha \lambda \sum_{i \neq j}^{mp} |\Omega_{ij}| + (1 - \alpha) \lambda \sum_{k \neq \ell}^{p} \|\mathbf{\Omega}^{(jk)}\|_F, \quad (6)$$

 $0 \le \alpha \le 1$ 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 $P_{sgl}(\mathbf{\Omega})$ is replaced with SCAD penalties, mimicking group SCAD of [15].

The first-order derivative of $P_{\lambda}(\theta)$ w.r.t. $|\theta|$ is $P'_{\lambda}(\theta) = \lambda$ 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 $P''_{\lambda}(\theta) = \frac{-1}{a-1}$ 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 $\{x(t)\}_{t=0}^{n-1}$, $x \in \mathbb{R}^{mp}$, regularization and penalty parameters λ , α and ρ_0 , tolerances τ_{abs} and τ_{rel} , variable penalty factor μ , maximum number of iterations i_{max} **Output:** estimated inverse covariance $\hat{\Omega}$ and edge-set $\hat{\mathcal{E}}$

- 1: Calculate sample covariance $\hat{\Sigma} = \frac{1}{n} \sum_{t=0}^{n-1} \boldsymbol{x}(t) \boldsymbol{x}^{\top}(t)$ (after centering $\boldsymbol{x}(t)$).
- 2: Initialize: $U^{(0)} = W^{(0)} = 0$, $\Omega^{(0)} = (\text{diag}(\hat{\Sigma}))^{-1}$, where $U, W \in \mathbb{R}^{(mp) \times (mp)}$, $\rho^{(0)} = \rho_0$
- 3: converged = false, i = 0
- 4: while converged = false and $i \leq i_{max}$, do
- 5: Eigen-decompose $\hat{\Sigma} \rho^{(i)} (\boldsymbol{W}^{(i)} \boldsymbol{U}^{(i)})$ as $\hat{\Sigma} \rho^{(i)} (\boldsymbol{W}^{(i)} \boldsymbol{U}^{(i)}) = \boldsymbol{V} \boldsymbol{D} \boldsymbol{V}^{\top}$ with diagonal matrix \boldsymbol{D} consisting of eigenvalues. Define diagonal matrix $\tilde{\boldsymbol{D}}$ with ℓ th diagonal element $\tilde{\boldsymbol{D}}_{\ell\ell} = (-\boldsymbol{D}_{\ell\ell} + \sqrt{\boldsymbol{D}_{\ell\ell}^2 + 4\rho^{(i)}})/(2\rho^{(i)})$. Set $\boldsymbol{\Omega}^{(i+1)} = \boldsymbol{V} \tilde{\boldsymbol{D}} \boldsymbol{V}^{\top}$. 6: Set $\boldsymbol{A}^{(jk)} = (\boldsymbol{\Omega}^{(jk)})^{(i+1)} + (\boldsymbol{U}^{(jk)})^{(i)}$. Define soft
- 6: Set $A^{(jk)} = (\Omega^{(jk)})^{(i+1)} + (U^{(jk)})^{(i)}$. Define soft thresholding scalar operator $S(a,\beta) := (1 - \beta/|a|)_{+}a$ where $(a)_{+} := \max(0, a)$. The diagonal $m \times m$ subblocks of W are updated as

$$[(\mathbf{W}^{(jj)})^{(i+1)}]_{st} = \begin{cases} [\mathbf{A}^{(jj)}]_{ss} & \text{if } s = t\\ S([\mathbf{A}^{(jj)}]_{st}, \frac{\alpha\lambda}{\rho^{(i)}}) & \text{if } s \neq t \end{cases}$$

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

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

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

8: Check convergence. Set tolerances

$$\tau_{pri} = mp \tau_{abs} + \tau_{rel} \max(\|\mathbf{\Omega}^{(i+1)}\|_F, \|\mathbf{W}^{(i+1)}\|_F)$$

$$\tau_{dual} = mp \tau_{abs} + \tau_{rel} \|\mathbf{U}^{(i+1)}\|_F / \rho^{(i)}.$$

Define $d_p = \|\mathbf{\Omega}^{(i+1)} - \mathbf{W}^{(i+1)}\|_F$ and $d_d = \rho^{(i)} \|\mathbf{W}^{(i+1)} - \mathbf{W}^{(i)}\|_F$. If $(d_p \leq \tau_{pri})$ and $(d_d \leq \tau_{dual})$, set converged = true.

9: Update penalty parameter ρ :

$$\rho^{(i+1)} = \begin{cases} 2\rho^{(i)} & \text{if } d_p > \mu d_d \\ \rho^{(i)}/2 & \text{if } d_d > \mu d_p \\ \rho^{(i)} & \text{otherwise} \,. \end{cases}$$

We also need to set $U^{(i+1)} = U^{(i+1)}/2$ for $d_p > \mu d_d$ and $U^{(i+1)} = 2U^{(i+1)}$ for $d_d > \mu d_p$.

10:
$$i \leftarrow i +$$

7:

11: end while

12: For $j \neq k$, if $\|\mathbf{W}^{(jk)}\|_F > 0$, assign edge $\{j, k\} \in \hat{\mathcal{E}}$, else $\{j, k\} \notin \hat{\mathcal{E}}$. Inverse covariance estimate $\hat{\mathbf{\Omega}} = \mathbf{W}$.

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

$$\min_{\boldsymbol{\Omega}\succ\boldsymbol{0},\boldsymbol{W}} \left\{ \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}) - \ln(|\boldsymbol{\Omega}|) + P_{sgl}(\boldsymbol{W}) \right\} \text{ subject to } \boldsymbol{\Omega} = \boldsymbol{W}$$

The scaled augmented Lagrangian for this problem is [21]

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

where 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 ρ following [21, Sec. 3.4.1]. At (i + 1)st iteration, the primal residual is given by $\Omega^{(i+1)} - W^{(i+1)}$ and the dual residual by $\rho^{(i)}(W^{(i+1)} - W^{(i)})$. Convergence criterion is met when the norms of these residuals are below tolerances τ_{pri} and τ_{dual} , respectively; see line 8 of Algorithm 1. In turn, τ_{pri} and τ_{dual} are chosen using an absolute and relative criterion as in line 8 of Algorithm 1 where τ_{abs} and τ_{rel} are user chosen absolute and relative tolerances, respectively.

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

$$P_{lin}(\boldsymbol{W}) = \sum_{i \neq j}^{mp} P'_{\alpha\lambda} (\hat{W}_{ij}^{(1)}) |W_{ij}| + \sum_{k \neq \ell}^{p} P'_{(1-\alpha)\lambda} (\|(\hat{\boldsymbol{W}}^{(1)})^{(k\ell)}\|_F) \|\boldsymbol{W}^{(jk)}\|_F.$$
(8)

Again solve a convex SGL problem after replacing $P_{sgl}(\boldsymbol{W})$ with $P_{lin}(\boldsymbol{W})$, and with following "obvious" modifications to Algorithm 1: in line 6 therein, replace $\alpha\lambda$ with $P'_{\alpha\lambda}(\hat{W}_{ij}^{(1)})$, and replace $(1 - \alpha)\lambda$ with $P'_{(1-\alpha)\lambda}(\|(\hat{\boldsymbol{W}}^{(1)})^{(k\ell)}\|_F)$. Recall that $P'_{\lambda}(\theta) = \lambda$ 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 Ω_0 denote the true Ω and \mathcal{E}_0 denote the true edgeset. Assume

(A1) $\operatorname{Card}(\mathcal{E}_0) = |\mathcal{E}_0| \leq s_{n0}.$

- (A2) $0 < \beta_{\min} \le \phi_{\min}(\Sigma_0) \le \phi_{\max}(\Sigma_0) \le \beta_{\max} < \infty$ where $\Sigma_0 = \Omega_0^{-1}$, and β_{\min} and β_{\max} are not functions of n.
- (A3) $\min_{\{\{i,j\}:\Omega_{0ij}\neq 0\}} |\Omega_{0ij}| \ge \delta_0 > 0.$

Let $\hat{\Omega}_{\lambda} = \arg \min_{\boldsymbol{\Omega} \succ \mathbf{0}} L(\boldsymbol{X}; \boldsymbol{\Omega})$. We denote p by p_n to indicate that it can grow with n.

Theorem 1 (Consistency): For $\tau > 2$, let

$$C_0 = 40 \, \max_k(\Sigma_{0kk}) \sqrt{2 \, (\tau + \ln(4) / \ln(mp_n))} \,. \tag{9}$$

For $\delta_1 \in (0,1)$ and "small" $\delta_2 > 0$, let

$$M = (1 + \delta_1)^2 (2 + \delta_2) C_0 / \beta_{\min}^2, \tag{10}$$

$$r_n = \sqrt{\frac{(mp_n + m^2 s_{n0}) \ln(mp_n)}{n}} = o(1), \qquad (11)$$

$$N_1 = 2\left(\ln(4) + \tau \,\ln(mp_n)\right),\tag{12}$$

$$N_2 = \arg\min\left\{n \, : \, r_n \le \frac{\delta_1 \beta_{\min}}{(1+\delta_1)^2 (2+\delta_2) C_0}\right\} \,.$$
(13)

Pick λ_n and integer N_3 as (a > 2 is a SCAD parameter)

$$\lambda_n = \begin{cases} \max(\frac{1}{\alpha}, \frac{1}{1-\alpha}) \max(M, C_0) r_n, & \alpha \in (0, 1) \\ \max(M, C_0) r_n, & \alpha = 0 \text{ or } 1 \end{cases}$$
(14)

$$N_3 = \arg\min\left\{n : \lambda_n < \frac{\min_{\{\{i,j\}:\Omega_{0ij}\neq 0\}} |\Omega_{0ij}|}{a}\right\}.$$
 (15)

For $n > \max\{N_1, N_2, N_3\}$, under assumptions (A1)-(A3), there exists a local minimizer $\hat{\Omega}_{\lambda}$ such that

$$\|\hat{\mathbf{\Omega}}_{\lambda} - \mathbf{\Omega}_0\|_F \le M r_n \tag{16}$$

with probability > $1 - 1/(mp_n)^{\tau-2}$. In terms of rate of convergence, $\|\hat{\Omega}_{\lambda} - \Omega_0\|_F = \mathcal{O}_P(r_n)$ •

V. PROOF OF THEOREM 1

Lemma 1 follows from [20, Lemma 1].

Lemma 1: Under Assumption (A2), the sample covariance $\hat{\Sigma}$ satisfies the tail bound

$$P\left(\max_{k,\ell} \left| [\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0]_{kl} \right| > C_0 \sqrt{\frac{\ln(mp_n)}{n}} \right) \le \frac{1}{(mp_n)^{\tau-2}}$$
(17)

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 $\Omega = \Omega_0 + \Delta$ with both Ω , $\Omega_0 > 0$, and $Q(\Omega) := L(X; \Omega) - L(X; \Omega_0)$. The estimate $\hat{\Omega}_{\lambda}$, denoted by $\hat{\Omega}$ hereafter suppressing dependence upon λ , minimizes $Q(\Omega)$, or equivalently, $\hat{\Delta} = \hat{\Omega} - \Omega_0$ minimizes $G(\Delta) := Q(\Omega_0 + \Delta)$. We will follow, for the most part, the method of proof of [19, Theorem 1] pertaining to lasso penalty. Consider the set

$$\Theta_n(M) := \left\{ \boldsymbol{\Delta} : \boldsymbol{\Delta} = \boldsymbol{\Delta}^\top, \| \boldsymbol{\Delta} \|_F = M r_n \right\}$$
(18)

where M and r_n are as in (10) and (11), respectively. Since $G(\hat{\Delta}) \leq G(\mathbf{0}) = 0$, if we can show that $\inf_{\Delta} \{G(\Delta) : \Delta \in \Theta_n(M)\} > 0$, then the minimizer $\hat{\Delta}$ must be inside $\Theta_n(M)$, and hence $\|\hat{\Delta}\|_F \leq Mr_n$. It is shown in [19, (9)] that $\ln(|\Omega_0 + \Delta|) - \ln(|\Omega_0|) = \operatorname{tr}(\Sigma_0 \Delta) - A_1$ where, with $H(\Omega_0, \Delta, v) = (\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1}$ and v denoting a scalar,

$$A_1 := \operatorname{vec}(\boldsymbol{\Delta})^\top \left(\int_0^1 (1-v) \boldsymbol{H}(\boldsymbol{\Omega}_0, \boldsymbol{\Delta}, v) \, dv \right) \operatorname{vec}(\boldsymbol{\Delta}) \,.$$
(19)

 α) λ_n , we can rewrite $G(\mathbf{\Delta})$ as

$$G(\mathbf{\Delta}) = \sum_{i=1}^{4} A_i, \quad A_2 := \operatorname{tr}\left((\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0)\mathbf{\Delta}\right)$$
(20)

$$A_{3} := \sum_{i \neq j}^{mp_{n}} (P_{\bar{\lambda}_{1}}(\Omega_{0ij} + \Delta_{ij}) - P_{\bar{\lambda}_{1}}(\Omega_{0ij})), \qquad (21)$$

$$A_4 := \sum_{k \neq \ell}^{p_n} (P_{\bar{\lambda}_2}(\|\mathbf{\Omega}_0^{(k\ell)} + \mathbf{\Delta}^{(k\ell)}\|_F) - P_{\bar{\lambda}_2}(\|\mathbf{\Omega}_0^{(k\ell)}\|_F))$$
(22)

Following [19, p. 502], we have

$$A_{1} \geq \frac{\|\mathbf{\Delta}\|_{F}^{2}}{2(\|\mathbf{\Omega}_{0}\| + \|\mathbf{\Delta}\|)^{2}} \geq \frac{\|\mathbf{\Delta}\|_{F}^{2}}{2\left(\beta_{\min}^{-1} + Mr_{n}\right)^{2}}$$
(23)

where we have used the fact that $\|\mathbf{\Omega}_0\| = \|\mathbf{\Sigma}_0^{-1}\| =$ $\phi_{\max}(\boldsymbol{\Sigma}_0^{-1}) = (\phi_{\min}(\boldsymbol{\Sigma}_0))^{-1} \leq \beta_{\min}^{-1} \text{ and } \|\boldsymbol{\Delta}\| \leq \|\boldsymbol{\Delta}\|_F =$ $Mr_n = \mathcal{O}(r_n)$. We now consider A_2 in (20). We have

$$A_2 = L_{21} + L_{22}, \ L_{22} = \sum_{\{i,j\}\in\bar{\mathcal{E}}_0^c} [\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0]_{ij} \Delta_{ji}, \quad (24)$$

$$L_{21} = \sum_{\{i,j\}\in\bar{\mathcal{E}}_0} [\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0]_{ij} \Delta_{ji} + \sum_i [\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0]_{ii} \Delta_{ii} \,. \tag{25}$$

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

$$|L_{21}| \leq \|\boldsymbol{\Delta}_{\bar{\mathcal{E}}_{0}}^{-} + \boldsymbol{\Delta}^{+}\|_{1} \max_{i,j} \left| [\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{0}]_{ij} \right| \\ \leq \sqrt{m^{2} s_{n0} + m p_{n}} \|\boldsymbol{\Delta}\|_{F} C_{0} \sqrt{\ln(m p_{n})/n} = C_{0} \|\boldsymbol{\Delta}\|_{F} r_{n}.$$
(26)

We consider L_{22} later as a part of A_3 where

$$A_3 = L_{31} + L_{32}, \ L_{32} = \sum_{\{i,j\} \in \bar{\mathcal{E}}_0^c} P_{\bar{\lambda}_1}(\Delta_{ij})$$
(27)

$$L_{31} = \sum_{\{i,j\}\in\bar{\mathcal{E}}_0} (P_{\bar{\lambda}_1}(\Omega_{0ij} + \Delta_{ij}) - P_{\bar{\lambda}_1}(\Omega_{0ij})).$$
(28)

For λ_n as in (14), $\overline{\lambda}_1 \ge Mr_n \ge |\Delta_{ij}|$ (since $\|\mathbf{\Delta}\|_F = Mr_n$), leading to $P_{\overline{\lambda}_1}(\Delta_{ij}) = \alpha \lambda_n |\Delta_{ij}|$. Consider L_{32} with αL_{22}

$$L_{32} - \alpha |L_{22}| \geq \sum_{\{i,j\} \in \bar{\mathcal{E}}_0^c} \left(\alpha \lambda_n |\Delta_{ij}| - |[\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0]_{ij}| |\Delta_{ij}| \right)$$
$$\geq \alpha \lambda_n \left(1 - \frac{C_0}{\alpha \lambda_n} \sqrt{\frac{\ln(mp_n)}{n}} \right) \sum_{\{i,j\} \in \bar{\mathcal{E}}_0^c} |\Delta_{ij}| > 0 \quad (29)$$

with prob. $> 1 - 1/(mp_n)^{\tau-2}$, since $\frac{C_0}{\alpha\lambda_n}\sqrt{\ln(mp_n)/n} < 1$. Now we bound $|L_{31}|$. A Taylor series expansion of $P_{\lambda}(\theta)$ for $\theta > 0$, around $\theta_0 > 0$, is given by $P_{\lambda}(\theta) = P_{\lambda}(\theta_0) +$ $P'_{\lambda}(\theta_0)(\theta - \theta_0) + P''_{\lambda}(\tilde{\theta})\frac{(\theta - \theta_0)^2}{2} \text{ where } \tilde{\theta} = \theta_0 + \gamma(\theta - \theta_0)$ for some $\gamma \in [0, 1]$. Setting $\lambda = \bar{\lambda}_1, \theta_0 = |\Omega_{0ij}|$ and $\theta =$ $|\Omega_{0ij} + \Delta_{ij}|$, and noting that $P_{\lambda}''(\tilde{\theta}) \leq 0$ for any $\tilde{\theta} > 0$, and $|\Omega_{0ij}| > 0$ for $\{i, j\} \in \mathcal{E}_0$, we have $P_{\overline{\lambda}_1}(\Omega_{0ij} + \Delta_{ij}) \leq 1$

Noting that $\Omega^{-1} = \Sigma$ and setting $\bar{\lambda}_1 = \alpha \lambda_n$ and $\bar{\lambda}_2 = (1 - P_{\bar{\lambda}_1}(\Omega_{0ij}) + P'_{\bar{\lambda}_1}(\Omega_{0ij})(|\Omega_{0ij} + \Delta_{ij}| - |\Omega_{0ij}|)$. Since $P'_{\bar{\lambda}_1}(\theta) \ge P'_{\bar{\lambda}_1}(\theta)$ $0 \ \forall \theta, P'_{\overline{\lambda}_1}(|\Omega_{0ij}|) = 0 \text{ for } n \geq N_3, \{i, j\} \in \overline{\mathcal{E}}_0,$

$$|L_{31}| \leq \sum_{\{i,j\}\in\bar{\mathcal{E}}_0} P'_{\bar{\lambda}_1}(|\Omega_{0ij}|) ||\Omega_{0ij} + \Delta_{ij}| - |\Omega_{0ij}||$$

=0 for $n \geq N_3$. (30)

Now consider A_4 which can be expressed as

$$A_{4} = L_{41} + L_{42}, \quad L_{42} = \sum_{\{k,\ell\}\in\mathcal{E}_{0}^{c}} P_{\bar{\lambda}_{2}}(\mathbf{\Delta}^{(k\ell)} \|_{F}), \quad (31)$$
$$L_{41} = \sum_{\{k,\ell\}\in\mathcal{E}_{0}} (P_{\bar{\lambda}_{2}}(\|\mathbf{\Omega}_{0}^{(k\ell)} + \mathbf{\Delta}^{(k\ell)}\|_{F}) - P_{\bar{\lambda}_{1}}(\|\mathbf{\Omega}_{0}^{(k\ell)}\|_{F})).$$

Similar to $|L_{31}|$, we have

$$|L_{41}| \leq \sum_{\{k,\ell\} \in \mathcal{E}_0} P'_{\bar{\lambda}_2}(\|\mathbf{\Omega}_0^{(k\ell)}\|_F) \Big| \|\mathbf{\Omega}_0^{(k\ell)} + \mathbf{\Delta}^{(k\ell)}\|_F - \|\mathbf{\Omega}_0^{(k\ell)}\|_F \Big| = 0 \text{ for } n \geq N_3$$
(32)

since $P'_{\overline{\lambda}_2}(\|\mathbf{\Omega}_0^{(k\ell)}\|_F) = 0$ for $n \ge N_3$ if $\{k, \ell\} \in \mathcal{E}_0$ and since $\min_{k,\ell} \| \widehat{\mathbf{\Omega}}_0^{(k\ell)} \|_F \ge \min_{i,j} |\Omega_{0ij}|.$ Now consider L_{42} with (1 - 1) α)L₂₂. With u = (k-1)m + s and v = (l-1)m + t, we have

$$L_{42} - (1 - \alpha)|L_{22}| \geq \sum_{\{k,\ell\}\in\mathcal{E}_0^c} \left((1 - \alpha)\lambda_n \| \boldsymbol{\Delta}^{(k\ell)} \|_F - (1 - \alpha) \sum_{s,t=1}^m \| [\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0]_{uv} \| |\Delta_{uv}| \right)$$

$$\geq (1 - \alpha) \sum_{\{k,\ell\}\in\mathcal{E}_0^c} (\lambda_n \| \boldsymbol{\Delta}^{(k\ell)} \|_F - mC_0 \sqrt{\frac{\ln(mp_n)}{n}} \| \boldsymbol{\Delta}^{(k\ell)} \|_F)$$

$$\geq (1 - \alpha)\lambda_n \left(1 - \frac{mC_0}{(1 - \alpha)\lambda_n} \sqrt{\frac{\ln(mp_n)}{n}} \right) \sum_{\{k,\ell\}\in\mathcal{E}_0^c} \| \boldsymbol{\Delta}^{(k\ell)} \|_F$$
(33)

with prob. $> 1 - 1/(mp_n)^{\tau-2}$, since $\frac{mC_0}{(1-\alpha)\lambda_n}\sqrt{\ln(mp_n)/n} <$ 1. Combining A_2 , A_3 and A_4 , we have

$$A_{2}+A_{3}+A_{4} = \sum_{i=1}^{3} \sum_{j=1}^{2} L_{ij} \ge -|L_{21}| + L_{32} - \alpha |L_{22}| + L_{31} + L_{42} - (1 - \alpha) |L_{22}| + L_{41} \ge -|L_{21}| + L_{31} + L_{41} \ge C_{0} \|\mathbf{\Delta}\|_{F} r_{n} \text{ for } n \ge N_{3}$$
(34)

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 $\|\mathbf{\Delta}\|_F = Mr_n$, we have with probability $> 1 - 1/(mp_n)^{\tau-2}$,

$$G(\mathbf{\Delta}) \ge \|\mathbf{\Delta}\|_{F}^{2} \left[\frac{1}{2(\beta_{\min}^{-1} + Mr_{n})^{2}} - \frac{C_{0}}{M}\right].$$
 (35)

For $n \ge N_2$, if we pick M as specified in (10), we obtain $Mr_n \leq Mr_{N_2} \leq \delta_1/\beta_{\min}$. Then

$$\frac{1}{2(\beta_{\min}^{-1} + Mr_n)^2} \ge \frac{\beta_{\min}^2}{2(1+\delta_1)^2} = \frac{(2+\delta_2)C_0}{2M} > \frac{C_0}{M},$$

implying $G(\Delta) > 0$. For $\alpha = 0$, omit A_3 , and for $\alpha = 1$, omit A_4 from $G(\Delta)$, to get $G(\Delta) > 0$, completing the proof.

VI. SIMULATION EXAMPLE

Now we consider an Erdös-Rényi graph where p nodes are connected to each other with probability $p_{er} = 0.05$. In the upper triangular $\overline{\Omega}$, using the notation of (2), we set $[\overline{\Omega}^{(jk)}]_{st} = 0.5^{|s-t|}$ for $j = k = 1, \dots, p$, $s, t = 1, \dots, m$. For $j \neq k$, if the two nodes are not connected, we have $\overline{\Omega}^{(jk)} = \mathbf{0}$, and if nodes j and k are connected in the chain graph, then $[\overline{\Omega}^{(jk)}]_{st}$ is uniformly distributed over $[-0.4, -0.1] \cup [0.1, 0.4]$ if $s \neq t$, and $[\overline{\Omega}^{(jk)}]_{st} = 0$ if s = t. Now add $\gamma \mathbf{I}$ to Ω with γ picked to make minimum eigenvalue of $\Omega + \gamma \mathbf{I}$ equal to 0.5. With $\Phi \Phi^{\top} = (\Omega + \gamma \mathbf{I})^{-1}$, we generate $\mathbf{x} = \Phi \mathbf{w}$ with $\mathbf{w} \in \mathbb{R}^{mp}$ as Gaussian $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We generate n i.i.d. observations from \mathbf{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 $\|\hat{W} - \Omega_0\|_F / \|\Omega_0\|_F$ and corresponding F_1 values; p = 400.

We used the solution outlined in Sec. III with a = 3.7, $\rho_0 = 2, \ \mu = 10, \ \text{and} \ \tau_{abs} = \tau_{rel} = 10^{-4}.$ 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 = $\|\hat{W} - \Omega_0\|_F / \|\Omega_0\|_F$ where $\hat{W} = \hat{W}^{(2)}$ for SG-SCAD penalty and $\hat{W} = \hat{W}^{(1)}$ for SGL penalty. We first selected the tuning parameters (λ, α) 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 λ as $\lambda_n \propto \sqrt{s_{n0} + (p_n/m)}$. $m \cdot \sqrt{\ln(mp_n)/n}$ for SG-SCAD based on (11) and (14), and as $\lambda_n \propto m \cdot \sqrt{\ln(mp_n)/n}$ 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 Ω 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

- 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. 4254-4278, 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 ℓ₁-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.