On Graphical Modeling of High-Dimensional Long-Range Dependent Time Series

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 a high-dimensional stationary, multivariate long-range dependent (LRD) Gaussian time series. In a time series graph, each component of the vector series is represented by a distinct node, and associations between components are represented by edges between the corresponding nodes. In a recent work on graphical modeling of short-range dependent (SRD) Gaussian time series, the problem was cast as one of multi-attribute graph estimation for random vectors where a vector is associated with each node of the graph. At each node, the associated random vector consists of a time series component and its delayed copies. A theoretical analysis based on short-range dependence has been given in Tugnait (2022 ICASSP). In this paper we analyze this approach for LRD Gaussian time series and provide consistency results regarding convergence in the Frobenius norm of the inverse covariance matrix associated with the multi-attribute graph.**

I. Introduction

Graphical models are an important and useful tool for analyzing multivariate data [1]. Given a collection of random variables, one wishes to assess the relationship between two variables, conditioned on the remaining variables. Consider a graph $\mathcal{G} = (V, \mathcal{E})$ with a set of p vertices (nodes) $V = \{1, 2, \cdots, p\} = [p]$, and a corresponding set of (undirected) edges $\mathcal{E} \subseteq [p] \times [p]$. Also consider a stationary, zero-mean, $p-$dimensional multivariate Gaussian time series $\boldsymbol{x}(t)$, $t = 0, \pm 1, \pm 2, \cdots$, with ith component $x_i(t)$. Given $\{\boldsymbol{x}(t)\}$, in the corresponding graph $\mathcal{G}$, each component series $\{x_i(t)\}$ is represented by a node (i in V), and associations between components $\{x_i(t)\}$ and $\{x_j(t)\}$ are represented by edges between nodes i and j of $\mathcal{G}$. In a conditional independence graph (CIG), there is no edge between nodes i and j if and only if (iff) $x_i(t)$ and $x_j(t)$ are conditionally independent given the remaining p-2 scalar series $x_\ell(t)$, $\ell \in [p]$, $\ell \neq i$, $\ell \neq j$ [2].

Graphical models were originally developed for random vectors [3, p. 234]. Such models have been extensively studied, and found to be useful in a wide variety of applications [4], [5]. Graphical modeling of real-valued time-dependent data (stationary time series) originated with [6], followed by [2]. A key insight in [2] was to transform the series to the frequency domain and express the graph relationships in the frequency domain. Nonparametric approaches for graphical modeling of real time series in high-dimensional settings have been formulated in the form of penalized log-likelihood in frequency-domain in [7]–[10]. Recently in [11] (Tugnait, 2022 ICASSP), a time-domain approach to graph structure estimation for stationary Gaussian multivariate time series was presented. All these works [7]–[11] assume that dependent time series exhibits short-range dependence. A (zero-mean) univariate stationary time series $\{y(t)\}$ is said to have long-range dependence (long-memory) if its autocorrelation function $r_y(\tau) = E\{y(t+\tau)y(t)\}$ satisfies $\sum_{\tau=-\infty}^{\infty} |r_y(\tau)| = \infty$, and short-range dependence (short-memory) otherwise [12]. A (zero-mean) multivariate stationary time series $\{\boldsymbol{x}(t)\}$ (p-dimensional) is said to have long-range dependence (long-memory) if its autocorrelation function $\boldsymbol{R}_x(\tau) = E\{\boldsymbol{x}(t+\tau)\boldsymbol{x}^\top(t)\}$ has the structure $[\boldsymbol{R}_x(\tau)]_{jk} = [\tilde{R}(\tau)]_{jk}\tau^{d_j+d_k-1}$ where $[\boldsymbol{A}]_{jk}$ denotes (j,k)th component of $\boldsymbol{A}$, $d_j \in (0, 0.5)$, $j = 1, 2, \cdots, p$, and $\tilde{R}(\tau) \in \mathbb{R}^{p\times p}$ satisfies $\lim_{\tau\to\infty} \tilde{R}(\tau) = \bar{R}$ [13]. Otherwise, $\{\boldsymbol{x}(t)\}$ (p-dimensional) is said to have short-range dependence.

In [11], graphical modeling of time series was cast as one of multi-attribute graph estimation for random vectors where a vector is associated with each node of the graph. At each node, the associated random vector consists of a time series component and its delayed copies. A theoretical analysis based on short-range dependence has been given in [11]. In this paper we analyze the approach of [11] for LRD Gaussian time series and provide consistency results regarding convergence in the Frobenius norm of the inverse covariance matrix associated with the multi-attribute graph.

Notation: We use $\boldsymbol{S} \succeq 0$ and $\boldsymbol{S} \succ 0$ to denote that the symmetric matrix $\boldsymbol{S}$ is positive semi-definite and positive definite, respectively. For a set V, $|V|$ or card(V) denotes its cardinality. $\mathbb{Z}$ is the set of integers. Given $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, we use $\phi_{\min}(\boldsymbol{A})$, $\phi_{\max}(\boldsymbol{A})$, $|\boldsymbol{A}|$ and 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 define $\|\boldsymbol{B}\| = \sqrt{\phi_{\max}(\boldsymbol{B}^\top\boldsymbol{B})}$, $\|\boldsymbol{B}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{B}^\top\boldsymbol{B})}$ and $\|\boldsymbol{B}\|_1 = \sum_{i,j}|B_{ij}|$, where B_{ij} is the (i,j)-th element of $\boldsymbol{B}$ (also denoted by $[\boldsymbol{B}]_{ij}$). Given $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, $\boldsymbol{A}^+ = \mathrm{diag}(\boldsymbol{A})$ 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.

II. Multi-Attribute Formulation for Time Series Graphical Modeling [11]

Consider stationary Gaussian time series $\boldsymbol{x}(t) \in \mathbb{R}^p$, $t \in \mathbb{Z}$, with $E\{\boldsymbol{x}(t)\} = 0$ and $\boldsymbol{R}_x(\tau) = \mathbb{E}\{\boldsymbol{x}(t+\tau)\boldsymbol{x}^T(t)\}$, $\tau \in \mathbb{Z}$.

This work was supported by NSF Grant ECCS-2040536. Author's email: tugnajk@auburn.edu

The conditional independence relationships among time series components $\{x_i(t)\}$'s are encoded in edge set $\mathcal{E}$ of $\mathcal{G}=(V,\mathcal{E})$, $V=[p]$, $\mathcal{E}\subseteq V\times V$, where edge $\{i,j\}\in\mathcal{E}$ iff $\{x_i(t),\, t\in\mathbb{Z}\}$ and $\{x_j(t),\, t\in\mathbb{Z}\}$ are conditionally independent given the remaining p-2 components $\boldsymbol{x}_{-ij,\mathbb{Z}}=\{x_k(t)\,:\,k\in V\backslash\{i,j\},\, t\in\mathbb{Z}\}$. Define $e_{i|-ij}(t)=x_i(t)-E\{x_i(t)|\boldsymbol{x}_{-ij,\mathbb{Z}}\}$, $e_{j|-ij}(t)=x_j(t)-E\{x_j(t)|\boldsymbol{x}_{-ij,\mathbb{Z}}\}$ and the power spectral density (PSD) matrix $\boldsymbol{S}_x(f)$, $\boldsymbol{S}_x(f)=\sum_{\tau=-\infty}^{\infty}\boldsymbol{R}_x(\tau)e^{-j2\pi f\tau}$. Then we have the following equivalence [2]

$$\begin{aligned}\text{edge } \{i,j\}\notin\mathcal{E} \;\Leftrightarrow\; & [\boldsymbol{S}_x^{-1}(f)]_{ij}=0\;\forall f\in[0,1]\\ \Leftrightarrow\; & E\{e_{i|-ij}(t+\tau)e_{j|-ij}(t)\}=0\;\forall\tau\in\mathbb{Z}\,. \end{aligned}\tag{1}$$

For some $d\geq 1$, let

$$\boldsymbol{z}_i(t)=[x_i(t)\; x_i(t-1)\;\cdots\; x_i(t-d)]^\top\in\mathbb{R}^{d+1}\tag{2}$$

$$\boldsymbol{y}(t)=[\boldsymbol{z}_1^\top(t)\;\boldsymbol{z}_2^\top(t)\;\cdots\;\boldsymbol{z}_p^\top(t)]^\top\in\mathbb{R}^{(d+1)p}\,.\tag{3}$$

We associate $\boldsymbol{z}_i$ with the ith node of graph $\mathcal{G}=(V,\mathcal{E})$, $V=[p]$, $\mathcal{E}\subseteq V\times V$. We now have $m=d+1$ attributes per node. Now $\{i,j\}\in\mathcal{E}$ iff vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are conditionally independent given the remaining p-2 vectors $\{\boldsymbol{z}_\ell\,,\ell\in V\backslash\{i,j\}\}$. Let $\boldsymbol{\Omega}_y=(E\{\boldsymbol{y}(t)\boldsymbol{y}^\top(t)\})^{-1}$. Define the $m\times m$ subblock $\boldsymbol{\Omega}_y^{(ij)}$ of $\boldsymbol{\Omega}_y$ as

$$[\boldsymbol{\Omega}_y^{(ij)}]_{rs}=[\boldsymbol{\Omega}_y]_{(i-1)m+r,(j-1)m+s}\,,\; r,s=1,2,\cdots,m\,.\tag{4}$$

Let $\boldsymbol{z}_{-ij}(t)=\{\boldsymbol{z}_k(t)\,:\,k\in V\backslash\{i,j\}\}$, $\boldsymbol{e}_{i|-ij}(t)=\boldsymbol{z}_i(t)-E\{\boldsymbol{z}_i(t)|\boldsymbol{z}_{-ij}(t)\}$, and $\boldsymbol{e}_{j|-ij}(t)=\boldsymbol{z}_j(t)-E\{\boldsymbol{z}_j(t)|\boldsymbol{z}_{-ij}(t)\}$. Then by multi-attribute graphical modeling [16],

$$\{i,j\}\notin\mathcal{E}\;\Leftrightarrow\;\boldsymbol{\Omega}_y^{(ij)}=\boldsymbol{0}\,.\tag{5}$$

Define $\tilde{\boldsymbol{x}}_{-ij;t,d}=\{\boldsymbol{x}_k(s)\,:\,k\in V\backslash\{i,j\}\,,t-d\leq s\leq t\}$, $e_{xi|-ij}(t')=x_i(t')-E\{x_i(t')|\tilde{\boldsymbol{x}}_{-ij;t,d}\}$, and $e_{xj|-ij}(t')=x_j(t')-E\{x_j(t')|\tilde{\boldsymbol{x}}_{-ij;t,d}\}$. Notice that $e_{xi|-ij}(t')$ is an element of $\boldsymbol{e}_{i|-ij}(t)$ for any $t-d\leq t'\leq t$. As shown in [11], we have

$$\begin{aligned}\boldsymbol{\Omega}_y^{(ij)}=\boldsymbol{0}\;\Leftrightarrow\; & E\{e_{xi|-ij}(t_1)e_{xj|-ij}(t_2)\}=0,\\ & \text{for}\quad t-d\leq t_1,t_2\leq t.\end{aligned}\tag{6}$$

It follows from (6) that if we let $d\uparrow\infty$, then checking if $\boldsymbol{\Omega}_y^{(ij)}=\boldsymbol{0}$ to ascertain (5) becomes a surrogate for checking if the last equivalence in (1) holds true for time series graph structure estimation without using frequency-domain methods. This is the approach followed in [11].

III. Sparse-Group Graphical Lasso Solution

Consider a finite set of data comprised of n zero-mean observations $\boldsymbol{x}(t)$, $t=0,1,2,\cdots,n-1$. Pick $d>1$ and as in (3), construct $\boldsymbol{y}(t)$ for $t=d,d+1,\cdots,n-1$ with sample size $\bar{n}=n-d$. Define the sample covariance $\hat{\boldsymbol{\Sigma}}_y=\frac{1}{\bar{n}}\sum_{t=d}^{n-1}\boldsymbol{y}(t)\boldsymbol{y}^\top(t)$. If the vector sequence $\{\boldsymbol{y}(t)\}_{t=d}^{n-1}$ were i.i.d., the log-likelihood (up to some constants) would be given by $\ln(|\boldsymbol{\Omega}_y|)-\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_y\boldsymbol{\Omega}_y)$ [17]. In our case the sequence is not i.i.d., but as in [11], we will still use this expression as a pseudo log-likelihood and following [11], [17], consider the penalized negative pseudo log-likelihood

$$L_{SGL}(\boldsymbol{\Omega}_y)=-\ln(|\boldsymbol{\Omega}_y|)+\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_y\boldsymbol{\Omega}_y)+P(\boldsymbol{\Omega}_y),\tag{7}$$

$$P(\boldsymbol{\Omega}_y)=\alpha\lambda\,\|\boldsymbol{\Omega}_y^-\|_1+(1-\alpha)m\lambda\sum_{j\neq k}^{p}\|\boldsymbol{\Omega}_y^{(jk)}\|_F\,,\tag{8}$$

where $P(\boldsymbol{\Omega}_y)$ is a sparse-group lasso penalty [4], [17], with group lasso penalty $(1-\alpha)m\lambda\sum_{j\neq k}^{p}\|\boldsymbol{\Omega}_y^{(jk)}\|_F$, $\lambda>0$ and lasso penalty $\alpha\lambda\,\|\boldsymbol{\Omega}_y^-\|_1$, $\lambda>0$ is a tuning parameter, and $0\leq\alpha\leq 1$. The function $L_{SGL}(\boldsymbol{\Omega}_y)$ is strictly convex in $\boldsymbol{\Omega}_y\succ\boldsymbol{0}$. Unlike [11], we use $(1-\alpha)m\lambda$ in $P(\boldsymbol{\Omega}_y)$ ([11] uses $(1-\alpha)\lambda$); we follow [18] by scaling based on number of grouped variables.

Following [17], an alternating direction method of multipliers (ADMM) approach (with variable splitting [19]) is discussed in [11] to minimize $L_{SGL}(\boldsymbol{\Omega}_y)$ w.r.t. $\boldsymbol{\Omega}_y$.

IV. Theoretical Analysis

Here we analyze consistency (Theorem 1) by invoking some results from [17], as in [11], except that unlike [11], here we consider LRD Gaussian time series.

To quantify the dependence structure of $\{\boldsymbol{x}(t)\}$, we will follow [15].

(A0) Assume $\{\boldsymbol{x}(t)\}$ obeys $\boldsymbol{x}(t)=\sum_{i=0}^{\infty}\boldsymbol{A}_i\boldsymbol{e}(t-i)$ where $\{\boldsymbol{e}(t)\}$ is i.i.d., Gaussian, zero-mean with identity covariance, $\boldsymbol{e}(t)\in\mathbb{R}^p$, $\boldsymbol{A}_i\in\mathbb{R}^{p\times p}$, and

$$\max_{1\leq q\leq p}\sqrt{\sum_{k=1}^{p}([\boldsymbol{A}_i]_{qk})^2}\leq\frac{c_a}{(\max(1,i))^\gamma}\tag{9}$$

for all $i\geq 0$, some $c_a\in(0,\infty)$, and $\gamma>\frac{1}{2}$.

Since autocorrelation function $\boldsymbol{R}_x(\tau)=E\{\boldsymbol{x}(t+\tau)\boldsymbol{x}^\top(t)\}=\sum_{i=0}^{\infty}\boldsymbol{A}_{i+\tau}\boldsymbol{A}_i^\top$, if $|[\boldsymbol{A}_i]_{qk}|=\mathcal{O}(i^{d_q-1})$, $d_q<0.5$, $q,k=1,2,\cdots,p$, then by [13, Prop. 3.1], $[\boldsymbol{R}_x(\tau)]_{qk}=\mathcal{O}(\tau^{d_q+d_k-1})$ as $|\tau|\to\infty$. Setting $\gamma=1-\min_q d_q$, it then follows from [13, Prop. 3.1] that $[\boldsymbol{R}_x(\tau)]_{qk}=\mathcal{O}(\tau^{d_q+d_k-1})=\mathcal{O}(\tau^{1-2\gamma})$, $\tau>0$. For $\gamma>1$, $\boldsymbol{R}_x(\tau)$ is (absolutely) summable, hence, $\{\boldsymbol{x}(t)\}$ has short-range dependence (SRD). For $\gamma\in(\frac{1}{2},1)$, $\boldsymbol{R}_x(\tau)$ is not summable, implying LRD [12]–[15].

In the SRD case, for $\gamma>1$, Assumption (A0) is satisfied if $\boldsymbol{x}(t)$ is generated by an asymptotically stable vector ARMA (autoregressive moving average) model with distinct "poles," satisfying $\boldsymbol{x}(t)=-\sum_{i=1}^{q}\boldsymbol{\Phi}_i\boldsymbol{x}(t-i)+\sum_{i=0}^{r}\boldsymbol{\Psi}_i\boldsymbol{e}(t-i)$, because in that case $\|\boldsymbol{A}_i\|_F\leq a|\lambda_0|^i$ for some $0<a<\infty$ where $|\lambda_0|<1$ is the largest magnitude "pole" (root of $c(z):=\left|\boldsymbol{I}+\sum_{i=1}^{q}\boldsymbol{\Phi}_iz^{-i}\right|=0$) of the model. It can be shown that there exist $0<b<\infty$ and $1<\gamma<\infty$ such that $a|\lambda_0|^i\leq b\,i^{-\gamma}$ for $i\geq 1$, thereby satisfying assumption (A0).

By Assumption (A0), it follows that with $\boldsymbol{y}(t)$ as in (3), $\boldsymbol{y}(t)=\sum_{i=0}^{\infty}\boldsymbol{B}_i\bar{\boldsymbol{e}}(t-i)$, $\bar{\boldsymbol{e}}(t)\in\mathbb{R}^{mp}$ is i.i.d., Gaussian, zero-mean with identity covariance, $m = d+1$, $\boldsymbol{B}_i \in \mathbb{R}^{(mp)\times(mp)}$, for some $\boldsymbol{B}_i$'s such that

$$\max_{1\le q\le mp} \sqrt{\sum_{k=1}^{mp} ([\boldsymbol{B}_i]_{qk})^2} \le \frac{c_a}{(\max(1,i))^\gamma} \tag{10}$$

for all $i \ge 0$, with c_a and γ as in Assumption (A0). Now we first restate [15, Lemma VI.2, supplementary] as Lemma 1 (γ is called β in [15]) as applied to $\hat{\boldsymbol{\Sigma}}_y = \frac{1}{\bar{n}}\sum_{t=d}^{n-1} \boldsymbol{y}(t)\boldsymbol{y}^\top(t)$, $\bar{n} = n-d$, $\boldsymbol{\Sigma}_{y0} = E\{\boldsymbol{y}(t)\boldsymbol{y}^\top(t)\}$. (The cases $\gamma = 1$ and $\gamma = \frac{3}{4}$ are not in [15, Lemma VI.2, supplementary], but can be inferred from [15, Lemma VI.1, supplementary].) It is the basis of our Lemma 2 used in Theorem 1.

Lemma 1: Under Assumption (A0), the elements of the sample covariance $\hat{\boldsymbol{\Sigma}}_y$ satisfy the tail bound

$$P\left(\left|[\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_{y0}]_{kl}\right| \ge z\right) \le 2\exp(-C_u \min(\frac{z^2}{L_{n,\gamma}}, \frac{z}{J_{n,\gamma}})) \tag{11}$$

for every $z > 0$ and every k, l, where $C_u \in (0,\infty)$ is a constant that depends only on c_a in (9) and γ, and $(L_{n,\gamma}^{-1}, J_{n,\gamma}^{-1}) = (\bar{n}, \bar{n})$ for $\gamma > 1$, $(L_{n,\gamma}^{-1}, J_{n,\gamma}^{-1}) = (\bar{n}, \bar{n}/\ln^2(\bar{n}))$ for $\gamma = 1$, $(L_{n,\gamma}^{-1}, J_{n,\gamma}^{-1}) = (\bar{n}, \bar{n}^{2\gamma-1})$ for $1 > \gamma > \frac{3}{4}$, $(L_{n,\gamma}^{-1}, J_{n,\gamma}^{-1}) = (\bar{n}/\ln(\bar{n}), \bar{n}^{2\gamma-1})$ for $\gamma = \frac{3}{4}$, and $(L_{n,\gamma}^{-1}, J_{n,\gamma}^{-1}) = (\bar{n}^{4\gamma-2}, \bar{n}^{2\gamma-1})$ for $\frac{3}{4} > \gamma > \frac{1}{2}$. •

Constant C_u results from the application of the Hanson-Wright inequality [20].

In rest of this section we allow p and λ to be a functions of sample size n, denoted as p_n and λ_n, respectively. Lemma 1 is now exploited to derive Lemma 2.

Lemma 2: Under Assumption (A0), the sample covariance $\hat{\boldsymbol{\Sigma}}_y$ satisfies the tail bound

$$P\left(\max_{k,l}\left|[\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_{y0}]_{kl}\right| > z_{*n}\right) \le \frac{1}{(mp_n)^{\tau-2}} \tag{12}$$

for $\tau > 2$, under the following conditions on $z_{*n} > 0$ and sample size $\bar{n} = n-d$:

(a) For $\gamma > \frac{3}{4}$:

$$z_{*n} = \sqrt{\ln(2(mp_n)^\tau)/(nC_u)}\,, \tag{13}$$

and sample size satisfies $\bar{n} \ge N_{z,\gamma} = N_a$ if $\gamma \in (1,\infty)$, where

$$N_a = \ln(2(mp_n)^\tau)/C_u\,, \tag{14}$$

sample size satisfies $\bar{n} \ge N_{z,\gamma} = \arg\min\{\bar{n} : \sqrt{\bar{n}}/\ln^2(\bar{n}) \ge N_a\}$ if $\gamma = 1$, and $\bar{n} \ge N_{z,\gamma} = \arg\min\{\bar{n} : \bar{n}^{(4\gamma-3)/2} \ge N_a\}$ if $\gamma \in (\frac{3}{4}, 1)$.

(b) For $\gamma = \frac{3}{4}$:

$$z_{*n} = \sqrt{\ln(2(mp_n)^\tau)\,\ln(n)/(nC_u)}\,, \tag{15}$$

and sample size satisfies $\bar{n} \ge N_{z,\gamma} = \arg\min\{\bar{n} : \bar{n}^{(4\gamma-3)/2}\sqrt{\ln(\bar{n})} \ge N_a\}$.

(c) For $\gamma \in (\frac{1}{2}, \frac{3}{4})$:

$$z_{*n} = \begin{cases} \sqrt{\ln(2(mp_n)^\tau)/(n^{4\gamma-2}C_u)} & \text{if} \quad N_a \le 1 \\ \ln(2(mp_n)^\tau)/(n^{2\gamma-1}C_u) & \text{if} \quad N_a > 1. \end{cases} \tag{16}$$

Proof. For $\gamma \in (1,\infty)$ and $\bar{n} \ge N_a^2$, using Lemma 1 it is easy to verify that $z_{*n}^2/L_{n,\gamma} \le z_{*n}/J_{n,\gamma}$ for specified z_{*n}. Applying the union bound over all $(mp_n)^2$ entries of $\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_{y0}$ in Lemma 1, we have

$$\begin{aligned} &P\left(\max_{k,l}\left|[\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_{y0}]_{kl}\right| > z_{*n}\right) \le P_{tb} \\ &= 2(mp_n)^2\exp\left(-C_u\frac{\ln(2(mp_n)^\tau)}{nC_u}\,n\right) \\ &= 2(mp_n)^2\exp\left(\ln(2(mp_n)^\tau)^{-1}\right) = \frac{1}{(mp_n)^{\tau-2}}\,. \end{aligned} \tag{17}$$

For $\gamma \in (\frac{3}{4}, 1]$ and stated conditions on $\bar{n}$, we again have $z_{*n}^2/L_{n,\gamma} \le z_{*n}/J_{n,\gamma}$ for the specified z_{*n}, and therefore, the desired result follows in a manner similar to (17). For part (b), when $\gamma = \frac{3}{4}$, for the specified z_{*n} and stated condition on $\bar{n}$, we have $z_{*n}^2/L_{n,\gamma} \le z_{*n}/J_{n,\gamma}$ for the specified z_{*n}, and therefore, the desired result follows as before. For $\gamma \in (\frac{1}{2}, \frac{3}{4})$, for the specified z_{*n}, it is easy to verify that we have $z_{*n}^2/L_{n,\gamma} \le z_{*n}/J_{n,\gamma}$ if $N_a \le 1$, else we have $z_{*n}^2/L_{n,\gamma} > z_{*n}/J_{n,\gamma}$ if $N_a > 1$. For $N_a \le 1$, the desired results follows in a manner similar to (17). If $N_a > 1$, $\gamma \in (\frac{1}{2}, \frac{3}{4})$ and $z_{*n} = \ln(2(mp_n)^\tau)/(n^{2\gamma-1}C_u)$, then applying the union bound to left-side of (12), we have

$$\begin{aligned} &P\left(\max_{k,l}\left|[\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_{y0}]_{kl}\right| > z_{*n}\right) \le P_{tb} \\ &= 2(mp_n)^2\exp\left(-C_u\frac{\ln(2(mp_n)^\tau)}{(n^{2\gamma-1}C_u)}\,n^{2\gamma-1}\right) \\ &= 2(mp_n)^2\exp\left(\ln(2(mp_n)^\tau)^{-1}\right) = \frac{1}{(mp_n)^{\tau-2}}\,. \end{aligned} \tag{18}$$

This completes the proof. ■

Lemma 2 above replaces [17, Lemma 2] for dependency in observations. Further assume

(A1) Let $\boldsymbol{\Sigma}_{y0} = E\{\boldsymbol{y}(t)\boldsymbol{y}^\top(t)\} \succ \boldsymbol{0}$ denote the true covariance of $\boldsymbol{y}(t)$. Define $\mathcal{E}_{y0} = \{\{i,j\} : \boldsymbol{\Omega}_{y0}^{(ij)} \ne \boldsymbol{0},\ i \ne j\}$ where $\boldsymbol{\Omega}_{y0} = \boldsymbol{\Sigma}_{y0}^{-1}$. Assume that card$(\mathcal{E}_{y0}) = |\mathcal{E}_{y0}| \le s_{n0}$.

(A2) The minimum and maximum eigenvalues of $\boldsymbol{\Sigma}_{y0}$ satisfy

$$0 < \beta_{\min} \le \phi_{\min}(\boldsymbol{\Sigma_{y0}}) \le \phi_{\max}(\boldsymbol{\Sigma_{y0}}) \le \beta_{\max} < \infty\,.$$

Here $\beta_{\min}$ and $\beta_{\max}$ are not functions of n.

Let $\hat{\boldsymbol{\Omega}}_{y\lambda} = \arg\min_{\boldsymbol{\Omega}_y \succ \boldsymbol{0}} L_{SGL}(\boldsymbol{\Omega}_y)$. Theorem 1 establishes consistency of $\hat{\boldsymbol{\Omega}}_{y\lambda}$ and its proof given in the Appendix, closely follows the proof of [17, Theorem 1].

Theorem 1 (Consistency): Let $\tau > 2$, $m = d+1$, $\bar{n} = n-d$. and z_{*n} and $N_{z,\gamma}$ be as in Lemma 2. Given real numbers $\delta_1 \in (0,1)$, $\delta_2 > 0$ and $C_1 > 0$, let $C_2 = \sqrt{m} + 1 + C_1$, and

$$M = (1+\delta_1)^2(2C_2+\delta_2)/\beta_{\min}^2\,, \tag{19}$$

$$\tilde{z}_n := z_{*n}\sqrt{mp_n + m^2 s_{n0}} = o(1)\,, \tag{20}$$

$$N_b = \arg\min\left\{\bar{n} : \tilde{z}_n \le \frac{\delta_1\beta_{\min}}{(1+\delta_1)^2(2C_2+\delta_2)}\right\}\,. \tag{21}$$

Suppose the regularization parameter λ_n and $\alpha \in [0,1]$ satisfy

$$C_1\, z_{*n} \sqrt{1 + \frac{p_n}{m s_{n0}}} \geq \lambda_n \geq z_{*n}\,. \tag{22}$$

Then if assumptions (A0)-(A2) hold true, and the sample size $\bar{n} = n - d$ satisfies $\bar{n} > \max\{N_{z,\gamma}, N_b\}$ for $\gamma \geq \frac{3}{4}$ and $\bar{n} > N_b$ for $\gamma \in (\frac{1}{2}, \frac{3}{4})$, with probability $> 1 - 1/(mp_n)^{\tau-2}$, $\hat{\boldsymbol{\Omega}}_{y\lambda}$ satisfies

$$\|\hat{\boldsymbol{\Omega}}_{y\lambda} - \boldsymbol{\Omega}_{y0}\|_F \leq M \tilde{z}_n\,. \tag{23}$$

In terms of convergence rate, $\|\hat{\boldsymbol{\Omega}}_{y\lambda} - \boldsymbol{\Omega}_{y0}\|_F = \mathcal{O}_P\left(\tilde{z}_n\right)$ •

Notice that for $\gamma > \frac{3}{4}$, we have

$$\|\hat{\boldsymbol{\Omega}}_{y\lambda} - \boldsymbol{\Omega}_{y0}\|_F = \mathcal{O}_P(\sqrt{(mp_n + m^2 s_{n0})\ln(mp_n)/n}\,)$$

whereas for $\gamma \in (\frac{1}{2}, \frac{3}{4})$, we have

$$\|\hat{\boldsymbol{\Omega}}_{y\lambda} - \boldsymbol{\Omega}_{y0}\|_F = \mathcal{O}_P(\sqrt{(mp_n + m^2 s_{n0})\ln(mp_n)/n^{4\gamma-2}}\,)$$

which can be much slower (since $1 > 4\gamma - 2 > 0$). Also, we can allow m to increase with n so long as $\tilde{z}_n = o(1)$.

References

[1] S.L. Lauritzen, *Graphical Models*. Oxford, UK: Oxford Univ. Press, 1996.

[2] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, pp. 157-172, 2000.

[3] M. Eichler, "Graphical modelling of multivariate time series," *Probability Theory and Related Fields*, vol. 153, issue 1-2, pp. 233-268, June 2012.

[4] 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 (Methodological)*, vol. 76, pp. 373-397, 2014.

[5] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436-1462, 2006.

[6] D.R. Brillinger, "Remarks concerning graphical models of times series and point processes," *Revista de Econometria (Brazilian Rev. Econometr.)*, vol. 16, pp. 1-23, 1996.

[7] A. Jung, G. Hannak and N. Goertz, "Graphical LASSO based model selection for time series," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1781-1785, Oct. 2015.

[8] J.K. Tugnait, "Graphical modeling of high-dimensional time series," in *Proc. 52nd ACSSC*, Pacific Grove, CA, Oct. 29 - Oct. 31, 2018, pp. 840-844.

[9] J.K. Tugnait, "Consistency of sparse-group lasso graphical model selection for time series," in *Proc. 54th ACSSC*, Pacific Grove, CA, Nov. 1-4, 2020, pp. 589-593.

[10] J.K. Tugnait, "New results on graphical modeling of high-dimensional dependent time series," in *Proc. 55th ACSSC*, Pacific Grove, CA, Oct. 31 - Nov. 3, 2021.

[11] J.K. Tugnait, "Graph learning from multivariate dependent time series via a multi-attribute formulation" in *Proc. 2022 IEEE Intern. Conf. Acoustics, Speech & Signal Processing (ICASSP 2022)*, Singapore, May 22-27, 2022.

[12] W. Palma, *Long-Memory Time Series: Theory and Methods*. Hoboken, NJ: Wiley-Interscience, 2005.

[13] S. Kechagias and V. Pipiras, "Definitions and representations of multivariate long-range dependent time series," *Journal of Time Series Analysis*, Vol. 36, Issue 1, pp. 1-25, Jan. 2015.

[14] V. Pipiras and M. Taqqu, *Long-Range Dependence and Self-Similarity*. Cambridge: Cambridge Univ. Press, 2017.

[15] X. Chen, M. Xu and W.B. Wu, "Regularized estimation of linear functionals of precision matrices for high-dimensional time series," *IEEE Trans. Signal Process.*, vol. 64, no. 24, pp. 6459-6470, Dec. 15, 2016.

[16] M. Kolar, H. Liu and E.P. Xing, "Graph estimation from multi-attribute data," *J. Machine Learning Research*, vol. 15, pp. 1713-1750, 2014.

[17] J.K. Tugnait, "Sparse-group lasso for graph learning from multi-attribute data," *IEEE Trans. Signal Process.*, vol. 69, pp. 1771-1786, 2021. (Corrections, vol. 69, p. 4758, 2021.)

[18] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Society, Series B (Methodological)*, vol. 68, pp. 49-67, 2006.

[19] 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.

[20] M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-gaussian concentration," *Electronic Communications Probability*, vol. 18, no. 82, pp. 1-9, 2013.

[21] 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.

Appendix

Proof of Theorem 1: Let $\boldsymbol{\Omega}_y = \boldsymbol{\Omega}_{y0} + \boldsymbol{\Delta}$ with both $\boldsymbol{\Omega}_y$, $\boldsymbol{\Omega}_{y0} \succ \mathbf{0}$, and

$$Q(\boldsymbol{\Omega}) := L_{SGL}(\boldsymbol{\Omega}_{y0}) - L_{SGL}(\boldsymbol{\Omega}_{y0})\,. \tag{24}$$

The estimate $\hat{\boldsymbol{\Omega}}_{y\lambda}$, denoted by $\hat{\boldsymbol{\Omega}}$ hereafter suppressing dependence upon λ and y, minimizes $Q(\boldsymbol{\Omega})$, or equivalently, $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_{y0}$ minimizes $G(\boldsymbol{\Delta}) := Q(\boldsymbol{\Omega}_{y0} + \boldsymbol{\Delta})$. We will follow the method of proof of [17, Theorem 1], based on [21, Theorem 1]. Consider the set

$$\Theta_n(M) := \left\{ \boldsymbol{\Delta} \,:\, \boldsymbol{\Delta} = \boldsymbol{\Delta}^\top,\ \|\boldsymbol{\Delta}\|_F = M\tilde{z}_n \right\} \tag{25}$$

where M and $\tilde{z}_n$ are as in (19) and (20), respectively. Since $G(\hat{\boldsymbol{\Delta}}) \leq G(\mathbf{0}) = 0$, if we can show that $\inf_{\boldsymbol{\Delta}}\{G(\boldsymbol{\Delta}) \,:\, \boldsymbol{\Delta} \in \Theta_n(M)\} > 0$, then the minimizer $\hat{\boldsymbol{\Delta}}$ must be inside $\Theta_n(M)$, and hence $\|\hat{\boldsymbol{\Delta}}\|_F \leq M\tilde{z}_n$. It is shown in [21, (9)] that

$$\ln(|\boldsymbol{\Omega}_{\boldsymbol{y0}} + \boldsymbol{\Delta}|) - \ln(|\boldsymbol{\Omega}_{y0}|) = \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{y0}}\boldsymbol{\Delta}) - A_1 \tag{26}$$

where, with $\boldsymbol{H}(\boldsymbol{\Omega}_{y0}, \boldsymbol{\Delta}, v) = (\boldsymbol{\Omega}_{\boldsymbol{y0}} + v\boldsymbol{\Delta})^{-1} \otimes (\boldsymbol{\Omega}_{\boldsymbol{y0}} + v\boldsymbol{\Delta})^{-1}$ and v denoting a scalar,

$$A_1 = \mathrm{vec}(\boldsymbol{\Delta})^\top \left(\int_0^1 (1-v)\boldsymbol{H}(\boldsymbol{\Omega}_{y0}, \boldsymbol{\Delta}, v)\, dv \right) \mathrm{vec}(\boldsymbol{\Delta})\,. \tag{27}$$

Noting that $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}$ and setting $\bar{\lambda}_1 = \alpha\lambda_n$ and $\bar{\lambda}_2 = (1-\alpha)m\lambda_n$, we can rewrite $G(\boldsymbol{\Delta})$ as

$$G(\boldsymbol{\Delta}) = A_1 + A_2 + A_3 + A_4\,, \tag{28}$$

where

$$A_2 = \mathrm{tr}\left((\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{y0})\boldsymbol{\Delta} \right)\,, \tag{29}$$

$$A_3 = \bar{\lambda}_1 \left(\|\boldsymbol{\Omega}_{y0}^- + \boldsymbol{\Delta}^-\|_1 - \|\boldsymbol{\Omega}_{y0}^-\|_1 \right)\,, \tag{30}$$

$$A_4 = \bar{\lambda}_2 \sum_{i,j=1; i\neq j}^{p_n} \left(\|\boldsymbol{\Omega}_{y0}^{(ij)} + \boldsymbol{\Delta}^{(ij)}\|_F - \|\boldsymbol{\Omega}_{y0}^{(ij)}\|_F \right)\,. \tag{31}$$

Following [21, p. 502], we have

$$A_1 \geq \frac{\|\mathbf{\Delta}\|_F^2}{2(\|\mathbf{\Omega}_{y0}\| + \|\mathbf{\Delta}\|)^2} \geq \frac{\|\mathbf{\Delta}\|_F^2}{2\left(\beta_{\min}^{-1} + M\tilde{z}_n\right)^2} \tag{32}$$

where we have used the fact that $\|\mathbf{\Omega}_{y0}\| = \|\mathbf{\Sigma}_{y0}^{-1}\| = \phi_{\max}(\mathbf{\Sigma}_{y0}^{-1}) = (\phi_{\min}(\mathbf{\Sigma}_{y0}))^{-1} \leq \beta_{\min}^{-1}$ and $\|\mathbf{\Delta}\| \leq \|\mathbf{\Delta}\|_F = M\tilde{z}_n$. We now consider A_2 in (29). We have

$$A_2 = \underbrace{\sum_{i,j=1;i\neq j}^{mp_n} [\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{y0}]_{ij}\Delta_{ji}}_{L_1} + \underbrace{\sum_{i=1}^{mp_n}[\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{y0}]_{ii}\Delta_{ii}}_{L_2} \tag{33}$$

To bound L_1, using Lemma 2, with probability $> 1 - 1/(mp_n)^{\tau-2}$,

$$|L_1| \leq \|\mathbf{\Delta}^-\|_1 \max_{i,j} \left|[\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{y0}]_{ij}\right| \leq \|\mathbf{\Delta}^-\|_1 \, z_{*n} \,. \tag{34}$$

Similarly, by Cauchy-Schwartz inequality, Lemma 2 and (20),

$$|L_2| \leq \|\mathbf{\Delta}^+\|_1 \, z_{*n} \leq z_{*n} \sqrt{mp_n} \, \|\mathbf{\Delta}^+\|_F \leq \|\mathbf{\Delta}^+\|_F \tilde{z}_n \,. \tag{35}$$

Therefore, with probability $> 1 - 1/(mp_n)^{\tau-2}$,

$$|A_2| \leq \|\mathbf{\Delta}^-\|_1 \, z_{*n} + \|\mathbf{\Delta}^+\|_F \tilde{z}_n \,. \tag{36}$$

We now derive a different bound on A_2. Define $\tilde{\mathbf{\Delta}} \in \mathbb{R}^{p_n \times p_n}$ with (i,j)-th element $\tilde{\Delta}_{ij} = \|\mathbf{\Delta}^{(ij)}\|_F$, where $\mathbf{\Delta}^{(ij)}$ is defined from $\mathbf{\Delta}$ similar to (4). By Cauchy-Schwartz inequality,

$$\begin{aligned} \|\mathbf{\Delta}^-\|_1 = \sum_{i,j=1;i\neq j}^{mp_n} |\Delta_{ij}| \leq m\|\tilde{\mathbf{\Delta}}^-\|_1 \\ + \underbrace{\left(\sum_{k=1}^{p_n} \|\mathbf{\Delta}^{(kk)}\|_1 - \|\mathbf{\Delta}^+\|_1\right)}_{=:B} . \end{aligned} \tag{37}$$

Then using $\sum_k \|\mathbf{\Delta}^{(kk)}\|_1 \leq m \sum_k \tilde{\Delta}_{kk} \leq m\sqrt{p_n} \, \|\tilde{\mathbf{\Delta}}^+\|_F$, we have

$$|L_2| + z_{*n} \, B \leq z_{*n} \, (\sum_{k=1}^{p_n} \|\mathbf{\Delta}^{(kk)}\|_1) \leq \|\tilde{\mathbf{\Delta}}^+\|_F \sqrt{m} \, \tilde{z}_n \,.$$

Therefore, an alternative bound is

$$|A_2| \leq m\|\tilde{\mathbf{\Delta}}^-\|_1 \, z_{*n} + \sqrt{m} \, \|\tilde{\mathbf{\Delta}}^+\|_F \tilde{z}_n \,. \tag{38}$$

We now bound A_3 in (30). Let $\bar{\mathcal{E}}_{y0}$ denote the true enlarged edge-set corresponding to $\mathcal{E}_{y0}$ when one interprets multi-attribute model as a single-attribute model. Let $\bar{\mathcal{E}}_{y0}^c$ denote its complement. For an index set $\boldsymbol{B}$ and a matrix $\boldsymbol{C} \in \mathbb{R}^{p_n \times p_n}$, we write $\boldsymbol{C}_{\boldsymbol{B}}$ to denote a matrix in $\mathbb{R}^{p_n \times p_n}$ such that $[\boldsymbol{C}_{\boldsymbol{B}}]_{ij} = C_{ij}$ if $(i,j) \in \boldsymbol{B}$, and $[\boldsymbol{C}_{\boldsymbol{B}}]_{ij} = 0$ if $(i,j) \notin \boldsymbol{B}$. Then, by definition, $\mathbf{\Delta}^- = \mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}} + \mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}^c}$, and $\|\mathbf{\Delta}^-\|_1 = \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1 + \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}^c}\|_1$. We have

$$\begin{aligned} A_3 &= \bar{\lambda}_1(\|\mathbf{\Omega}_{y0}^- + \mathbf{\Delta}^-\|_1 - \|\mathbf{\Omega}_{y0}^-\|_1) \\ &= \bar{\lambda}_1(\|\mathbf{\Omega}_{y0}^- + \mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1 + \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}^c}\|_1 - \|\mathbf{\Omega}_{y0}^-\|_1) \\ &\geq \bar{\lambda}_1(\|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}^c}\|_1 - \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1) \end{aligned} \tag{39}$$

where we have used the triangle inequality $\|\mathbf{\Omega}_{y0}^- + \mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1 \geq \|\mathbf{\Omega}_{y0}^-\|_1 - \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1$. Next we bound A_4 in (31). Considering the true edge-set $\mathcal{E}_{y0}$ for the multi-attribute graph, let $\mathcal{E}_{y0}^c$ denote its complement. If the edge $\{i,j\} \in \mathcal{E}_{y0}^c$, then $\mathbf{\Omega}_{y0}^{(ij)} = \mathbf{0}$, therefore, $\|\mathbf{\Omega}_{y0}^{(ij)} + \mathbf{\Delta}^{(ij)}\|_F - \|\mathbf{\Omega}_{y0}^{(ij)}\|_F = \|\mathbf{\Delta}^{(ij)}\|_F$. For $\{i,j\} \in \mathcal{E}_{y0}$, by the triangle inequality, $\|\mathbf{\Omega}_{y0}^{(ij)} + \mathbf{\Delta}^{(ij)}\|_F - \|\mathbf{\Omega}_{y0}^{(ij)}\|_F \geq -\|\mathbf{\Delta}^{(ij)}\|_F$. Thus

$$A_4 \geq \bar{\lambda}_2(\|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}^c}\|_1 - \|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}}\|_1) \,. \tag{40}$$

Split A_2 as $A_2 = \alpha A_2 + (1-\alpha)A_2$, apply bound (36) to αA_2 and (38) to $(1-\alpha)A_2$, use $\|\mathbf{\Delta}^-\|_1 = \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1 + \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}^c}\|_1$ and $\|\tilde{\mathbf{\Delta}}^-\|_1 = \|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}}\|_1 + \|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}^c}\|_1$ to yield

$$\begin{aligned} A_2 + A_3 + A_4 &\geq -|A_2| + \bar{\lambda}_1(\|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}^c}\|_1 - \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1) \\ &\quad + \bar{\lambda}_2(\|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}^c}\|_1 - \|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}}\|_1) \\ &\geq -(\alpha\|\mathbf{\Delta}^+\|_F + (1-\alpha)\sqrt{m}\,\|\tilde{\mathbf{\Delta}}^+\|_F)\tilde{z}_n \\ &\quad + \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}^c}\|_1(\bar{\lambda}_1 - \alpha z_{*n}) \\ &\quad + \|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}^c}\|_1(\bar{\lambda}_2 - (1-\alpha)m z_{*n}) \\ &\quad - \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1(\bar{\lambda}_1 + \alpha z_{*n}) - \|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}}\|_1(\bar{\lambda}_2 + (1-\alpha)m z_{*n}) \\ &\geq -(\alpha + (1-\alpha)\sqrt{m})\,\|\mathbf{\Delta}\|_F\tilde{z}_n - \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1(\bar{\lambda}_1 + \alpha z_{*n}) \\ &\quad - \|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}}\|_1(\bar{\lambda}_2 + (1-\alpha)m z_{*n}) \end{aligned} \tag{41}$$

where we used the fact that for λ_n as in (22), $\bar{\lambda}_1 - \alpha z_{*n} \geq 0$ and $\bar{\lambda}_2 - (1-\alpha)m z_{*n} \geq 0$, and $\|\mathbf{\Delta}^+\|_F \leq \|\mathbf{\Delta}\|_F$, $\|\tilde{\mathbf{\Delta}}^+\|_F \leq \|\mathbf{\Delta}\|_F$. By Cauchy-Schwartz inequality,

$$\|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_1 \leq \sqrt{m^2 s_{n0}}\,\|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_{y0}}\|_F \leq m\sqrt{s_{n0}}\,\|\mathbf{\Delta}\|_F \,, \tag{42}$$

$$\|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}}\|_1 \leq \sqrt{s_{n0}}\,\|\tilde{\mathbf{\Delta}}^-_{\mathcal{E}_{y0}}\|_F \leq \sqrt{s_{n0}}\,\|\tilde{\mathbf{\Delta}}\|_F = \sqrt{s_{n0}}\,\|\mathbf{\Delta}\|_F \,. \tag{43}$$

Using (41)-(43) and $\alpha_m := (\alpha + (1-\alpha)\sqrt{m})$, we have

$$\begin{aligned} A_2 + A_3 + A_4 &\geq -\Big[z_{*n}\left(1 + \alpha_m\sqrt{1 + p_n/(m s_{n0})}\right) + \bar{\lambda}_1 \\ &\qquad + (\bar{\lambda}_2/m)\Big] m\sqrt{s_{n0}}\,\|\mathbf{\Delta}\|_F \\ &\geq -\Big[(\sqrt{m}+1)\tilde{z}_n + (\bar{\lambda}_1 + \frac{\bar{\lambda}_2}{m})m\sqrt{s_{n0}}\Big]\,\|\mathbf{\Delta}\|_F \\ &\geq -C_2\tilde{z}_n\,\|\mathbf{\Delta}\|_F \end{aligned} \tag{44}$$

where we used the fact that for λ_n as in (22), $m\sqrt{s_{n0}}(\bar{\lambda}_1 + (\bar{\lambda}_2/m)) \leq C_1\tilde{z}_n$, and $\alpha_m \leq \sqrt{m}$. Using (28), the bound (32) on A_1 and (44) on $A_2 + A_3 + A_4$, and $\|\mathbf{\Delta}\|_F = M\tilde{z}_n$, we have with probability $> 1 - 1/(mp_n)^{\tau-2}$,

$$G(\mathbf{\Delta}) \geq \|\mathbf{\Delta}\|_F^2\left[\frac{1}{2(\beta_{\min}^{-1} + M\tilde{z}_n)^2} - C_2\frac{C_0}{M}\right] . \tag{45}$$

For $\bar{n} \geq N_b$, if we pick M as specified in (19), we obtain $M\tilde{z}_n \leq M r_{N_2} \leq \delta_1/\beta_{\min}$. Then

$$\frac{1}{2(\beta_{\min}^{-1} + M\tilde{z}_n)^2} \geq \frac{\beta_{\min}^2}{2(1+\delta_1)^2} = \frac{(2C_2 + \delta_2)C_0}{2M} > C_2\frac{C_0}{M} \,,$$

implying $G(\mathbf{\Delta}) > 0$. This proves the desired result. ■