ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746603

GRAPH LEARNING FROM MULTIVARIATE DEPENDENT TIME SERIES VIA A MULTI-ATTRIBUTE FORMULATION

Jitendra K. Tugnait

Department of Electrical & Computer Engineering
Auburn University, Auburn, AL 36849, USA

ABSTRACT

We consider the problem of inferring the conditional independence graph (CIG) of a high-dimensional stationary multivariate Gaussian time series. In a time series graph, each component of the vector series is represented by distinct node, and associations between components are represented by edges between the corresponding nodes. We formulate the problem 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. We present an alternating direction method of multipliers (ADMM) solution to minimize a sparse-group lasso penalized negative pseudo log-likelihood objective function to estimate the precision matrix of the random vector associated with the entire multi-attribute graph. The time series CIG is then inferred from the estimated precision matrix. A theoretical analysis is provided. Numerical results illustrate the proposed approach which outperforms existing frequency-domain approaches in correctly detecting the graph edges.

Keywords: Sparse graph learning; graph estimation; time series; undirected graph; multi-attribute graphs.

1. 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. In graphical models, graphs are used to display the conditional independence structure of the 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 (real-valued), 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 (whose statistics are estimated via multiple independent realizations) [3, p. 234]. Such models have been extensively studied, and found to be useful in a wide variety of applications [4–8]. Graphical modeling of real-valued time-dependent data (stationary time series) originated with [9], 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 (p is large and/or sample size n is of the order of p) have been formulated in the form of group-lasso penalized log-likelihood in frequency-domain in [10]. Sparse-group lasso penalized log-likelihood approach in frequency-domain has been considered in [11–13].

In this paper we investigate graph structure estimation for stationary Gaussian multivariate time series using a time-domain approach, unlike [10–12] who, as noted earlier, use a frequency-domain approach. After reviewing some graphical modeling background in Sec. 2, we first reformulate the problem in Sec. 3 as one of multi-attribute graph estimation for random vectors where a vector is associated with each node of the graph. Then in Sec. 4 we exploit the results of [14] to provide an alternating direction method of multipliers (ADMM) solution to minimize a sparse-group lasso penalized negative pseudo log-likelihood objective function for multi-attribute graph precision matrix estimation. A theoretical analysis is provided in Sec. 5. Numerical results in Sec. 6 illustrate the proposed approach.

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 $\|\boldsymbol{B}\| = \sqrt{\phi_{\max}(\boldsymbol{B}^\top \boldsymbol{B})}$, $\|\boldsymbol{B}\|_F = \sqrt{\text{tr}(\boldsymbol{B}^\top \boldsymbol{B})}$ 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 $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, $\boldsymbol{A}^+ = \text{diag}(\boldsymbol{A})$ 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.

2. GRAPHICAL MODELS

Here we provide some background material for graphical models for random vectors and for multivariate time series.

2.1. Random Vectors

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 $\mathcal{E} \subseteq V \times V$. Let $\boldsymbol{x} = [x_1\ x_2\ \cdots\ x_p]^\top \in \mathbb{R}^p$ denote a Gaussian random vector that is zero-mean with covariance $\boldsymbol{\Sigma} = E\{\boldsymbol{x}\boldsymbol{x}^\top\} \succ \boldsymbol{0}$. The conditional independence relationships among x_i's are encoded in $\mathcal{E}$ where edge $\{i, j\}$ between nodes i and j exists if and only if (iff) x_i and x_j are conditionally independent given the remaining p-2 variables x_ℓ, $\ell \in [p], \ell \neq i, \ell \neq j$. Let

$$\boldsymbol{x}_{-ij} = \{x_k : k \in V \backslash \{i,j\}\} \in \mathbb{R}^{p-2} \tag{1}$$

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

denote the vector $\boldsymbol{x}$ after deleting x_i and x_j from it. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ denote the precision matrix. Define

$$e_{i|-ij} = x_i - E\{x_i|\boldsymbol{x}_{-ij}\}\,,\;\; e_{j|-ij} = x_j - E\{x_j|\boldsymbol{x}_{-ij}\}\,. \quad (2)$$

Then we have the following equivalence [1]

$$\{i,j\} \notin \mathcal{E} \;\Leftrightarrow\; \Omega_{ij} = 0 \;\Leftrightarrow\; E\{e_{i|-ij}e_{j|-ij}\} = 0\,. \quad (3)$$

Note that $E\{x_i|\boldsymbol{x}_{-ij}\}$ is linear in $\boldsymbol{x}_{-ij}$ since $\boldsymbol{x}$ is zero-mean Gaussian, and furthermore it minimizes the conditional mean-square error

$$E\{x_i|\boldsymbol{x}_{-ij}\} = \arg\min_b E\{(x_i - b(\boldsymbol{x}_{-ij}))^2|\boldsymbol{x}_{-ij}\}\,. \quad (4)$$

Similar comments apply to $E\{x_j|\boldsymbol{x}_{-ij}\}$.

2.2. Multivariate Time Series

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

$$\boldsymbol{x}_{-ij,\mathbb{Z}} = \{x_k(t) \,:\, k \in V\backslash\{i,j\},\, t \in \mathbb{Z}\}\,. \quad (5)$$

Define

$$e_{i|-ij}(t) = x_i(t) - E\{x_i(t)|\boldsymbol{x}_{-ij,\mathbb{Z}}\} \quad (6)$$

$$e_{j|-ij}(t) = x_j(t) - E\{x_j(t)|\boldsymbol{x}_{-ij,\mathbb{Z}}\}\,, \quad (7)$$

and the power spectral density (PSD) matrix $\boldsymbol{S}_x(f)$

$$\boldsymbol{S}_x(f) = \sum_{\tau=-\infty}^{\infty} \boldsymbol{R}_{xx}(\tau)e^{-j2\pi f\tau}\,. \quad (8)$$

Then we have the following equivalence [2]

$$\begin{aligned}\{i,j\} \notin \mathcal{E} \;&\Leftrightarrow\; [\boldsymbol{S}_x^{-1}(f)]_{ij} = 0\;\forall f \in [0,1] \\ &\Leftrightarrow\; E\{e_{i|-ij}(t+\tau)e_{j|-ij}(t)\} = 0\;\forall \tau \in \mathbb{Z}\,. \end{aligned} \quad (9)$$

2.3. Multi-Attribute Graphical Models for Random Vectors

Now consider p jointly Gaussian vectors $\boldsymbol{z}_i \in \mathbb{R}^m$, $i \in [p]$. 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 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

$$\boldsymbol{x} = [\boldsymbol{z}_1^\top\;\boldsymbol{z}_2^\top\;\cdots\;\boldsymbol{z}_p^\top]^\top \in \mathbb{R}^{mp}\,. \quad (10)$$

Let $\boldsymbol{\Omega} = (E\{\boldsymbol{x}\boldsymbol{x}^\top\})^{-1}$ assuming $E\{\boldsymbol{x}\boldsymbol{x}^\top\} \succ \boldsymbol{0}$. Define the $m \times m$ subblock $\boldsymbol{\Omega}^{(ij)}$ of $\boldsymbol{\Omega}$ as

$$[\boldsymbol{\Omega}^{(ij)}]_{rs} = [\boldsymbol{\Omega}]_{(i-1)m+r,(j-1)m+s}\,,\; r,s = 1,2,\cdots,m\,. \quad (11)$$

Let

$$\boldsymbol{z}_{-ij} = \{\boldsymbol{z}_k \,:\, k \in V\backslash\{i,j\}\} \in \mathbb{R}^{m(p-2)} \quad (12)$$

denote the vector $\boldsymbol{x}$ in (10) after deleting vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ from it. Define

$$\boldsymbol{e}_{i|-ij} = \boldsymbol{z}_i - E\{\boldsymbol{z}_i|\boldsymbol{z}_{-ij}\}\,,\;\; \boldsymbol{e}_{j|-ij} = \boldsymbol{z}_j - E\{\boldsymbol{z}_j|\boldsymbol{z}_{-ij}\}\,. \quad (13)$$

Then we have the following equivalence [15]

$$\{i,j\} \notin \mathcal{E} \;\Leftrightarrow\; \boldsymbol{\Omega}^{(ij)} = \boldsymbol{0} \;\Leftrightarrow\; E\{\boldsymbol{e}_{i|-ij}\boldsymbol{e}_{j|-ij}^\top\} = \boldsymbol{0}\,, \quad (14)$$

where the first equivalence in (14) is given in [15, Sec. 2.1] and the second equivalence is given in [15, Appendix B.3].

3. MULTI-ATTRIBUTE FORMULATION FOR TIME SERIES GRAPHICAL MODELING

Consider time series $\{\boldsymbol{x}(t)\}$ as in Sec. 2.2. For some $d \geq 1$, let

$$\boldsymbol{z}_i(t) = [x_i(t)\; x_i(t-1)\; \cdots\; x_i(t-d)]^\top \in \mathbb{R}^{d+1} \quad (15)$$

$$\boldsymbol{y}(t) = [\boldsymbol{z}_1^\top(t)\; \boldsymbol{z}_2^\top(t)\; \cdots\; \boldsymbol{z}_p^\top(t)]^\top \in \mathbb{R}^{(d+1)p}\,. \quad (16)$$

Let $\boldsymbol{\Omega}_y = (E\{\boldsymbol{y}(t)\boldsymbol{y}^\top(t)\})^{-1}$. With $m = d+1$, define the $m \times m$ subblock $\boldsymbol{\Omega}_y^{(ij)}$ of $\boldsymbol{\Omega}_y$ as

$$[\boldsymbol{\Omega}_y^{(ij)}]_{rs} = [\boldsymbol{\Omega}_y]_{(i-1)m+r,(j-1)m+s}\,,\; s,t = 1,2,\cdots,m\,. \quad (17)$$

Let

$$\boldsymbol{z}_{-ij}(t) = \{\boldsymbol{z}_k(t) \,:\, k \in V\backslash\{i,j\}\}\,, \quad (18)$$

$$\boldsymbol{e}_{i|-ij}(t) = \boldsymbol{z}_i(t) - E\{\boldsymbol{z}_i(t)|\boldsymbol{z}_{-ij}(t)\} \quad (19)$$

$$\boldsymbol{e}_{j|-ij}(t) = \boldsymbol{z}_j(t) - E\{\boldsymbol{z}_j(t)|\boldsymbol{z}_{-ij}(t)\}\,. \quad (20)$$

Then by Sec. 2.3,

$$\{i,j\} \notin \mathcal{E} \;\Leftrightarrow\; \boldsymbol{\Omega}_y^{(ij)} = \boldsymbol{0}\,. \quad (21)$$

Define

$$\tilde{\boldsymbol{x}}_{-ij;t,d} = \{\boldsymbol{x}_k(s) \,:\, k \in V\backslash\{i,j\}\,, t-d \leq s \leq t\}\,, \quad (22)$$

$$e_{xi|-ij}(t') = x_i(t') - E\{x_i(t')|\tilde{\boldsymbol{x}}_{-ij;t,d}\} \quad (23)$$

$$e_{xj|-ij}(t') = x_j(t') - E\{x_j(t')|\tilde{\boldsymbol{x}}_{-ij;t,d}\}\,. \quad (24)$$

Notice that $e_{xi|-ij}(t')$ above is an element of $\boldsymbol{e}_{i|-ij}(t)$ defined in (19) for any $t-d \leq t' \leq t$. Then by (14) and (21), we have

$$\boldsymbol{\Omega}_y^{(ij)} = \boldsymbol{0} \;\Leftrightarrow\; E\{e_{xi|-ij}(t_1)e_{xj|-ij}(t_2)\} = 0,\;\; t-d \leq t_1, t_2 \leq t. \quad (25)$$

It follow from (25) that if we let $d \uparrow \infty$, then checking if $\boldsymbol{\Omega}_y^{(ij)} = \boldsymbol{0}$ to ascertain (21) becomes a surrogate for checking if the last equivalence in (9) holds true for time series graph structure estimation without using frequency-domain methods.

4. SPARSE-GROUP GRAPHICAL LASSO SOLUTION TO MULTI-ATTRIBUTE FORMULATION

We now 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 (16), 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 $\ln(|\boldsymbol{\Omega}_y|) - \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_y\boldsymbol{\Omega}_y)$ [14]. In our case the sequence is not i.i.d., but we will still use this expression as a pseudo log-likelihood and following [14], consider the penalized negative pseudo log-likelihood

$$L_{SGL}(\boldsymbol{\Omega}_y) = -\ln(|\boldsymbol{\Omega}_y|) + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_y\boldsymbol{\Omega}_y) + P(\boldsymbol{\Omega}_y), \quad (26)$$

$$P(\boldsymbol{\Omega}_y) = \alpha\lambda\,\|\boldsymbol{\Omega}_y^-\|_1 + (1-\alpha)\lambda\sum_{j\neq k}^{p}\|\boldsymbol{\Omega}_y^{(jk)}\|_F\,, \quad (27)$$

where $P(\boldsymbol{\Omega}_y)$ is a sparse-group lasso penalty [4, 14, 16, 17], with group lasso penalty $(1-\alpha)\lambda\sum_{j\neq k}^{p}\|\boldsymbol{\Omega}_y^{(jk)}\|_F$, $\lambda > 0$ and lasso penalty $\alpha\lambda\,\|\boldsymbol{\Omega}_y^-\|_1$, $\lambda > 0$ is a tuning parameter, and $0 \leq \alpha \leq 1$

yields a convex combination of lasso and group lasso penalties. The function $L_{SGL}(\boldsymbol{\Omega}_y)$ is strictly convex in $\boldsymbol{\Omega}_y \succ \mathbf{0}$.

As in [14], we use the ADMM approach [18] with variable splitting. Using variable splitting, consider

$$\min_{\boldsymbol{\Omega}_y \succ \mathbf{0}, \boldsymbol{W}} \Big\{ \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_y \boldsymbol{\Omega}_y) - \ln(|\boldsymbol{\Omega}_y|) + P(\boldsymbol{W}) \Big\} \text{ subject to } \boldsymbol{\Omega}_y = \boldsymbol{W} \,. \tag{28}$$

The scaled augmented Lagrangian for this problem is [18]

$$L_\rho = \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_y \boldsymbol{\Omega}_y) - \ln(|\boldsymbol{\Omega}_y|) + P(\boldsymbol{W}) + \frac{\rho}{2} \|\boldsymbol{\Omega}_y - \boldsymbol{W} + \boldsymbol{U}\|_F^2 \tag{29}$$

where $\boldsymbol{U}$ is the dual variable, and $\rho > 0$ is the penalty parameter. Given the results $\boldsymbol{\Omega}^{(i)}, \boldsymbol{W}^{(i)}, \boldsymbol{U}^{(i)}$ of the ith iteration, in the $(i+1)$st iteration, an ADMM algorithm executes the following three updates:

(a) $\boldsymbol{\Omega}_y^{(i+1)} \leftarrow \arg\min_{\boldsymbol{\Omega}_y} \; L_a(\boldsymbol{\Omega}_y), \;\; L_a(\boldsymbol{\Omega}_y) := \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_y \boldsymbol{\Omega}_y) - \ln(|\boldsymbol{\Omega}_y|) + \frac{\rho}{2} \|\boldsymbol{\Omega}_y - \boldsymbol{W}^{(i)} + \boldsymbol{U}^{(i)}\|_F^2$

(b) $\boldsymbol{W}^{(i+1)} \leftarrow \arg\min_{\boldsymbol{W}} L_b(\boldsymbol{W}), \;\; L_b(\boldsymbol{W}) := \alpha\lambda \|\boldsymbol{W}^-\|_1 + (1-\alpha)\lambda \sum_{i \neq j}^p \|\boldsymbol{W}^{(ij)}\|_F + \frac{\rho}{2} \|\boldsymbol{\Omega}_y^{(i+1)} - \boldsymbol{W} + \boldsymbol{U}^{(i)}\|_F^2$

(c) $\boldsymbol{U}^{(i+1)} \leftarrow \boldsymbol{U}^{(i)} + \Big(\boldsymbol{\Omega}_y^{(i+1)} - \boldsymbol{W}^{(i+1)} \Big)$

Remark 1. We follow the detailed ADMM algorithm given in [14] for the above updates; details may be found therein (where we need to replace $\boldsymbol{\Omega}$ with $\boldsymbol{\Omega}_y$). The parameter tuning (selection of λ and α) approach given in [14] does not apply (strictly speaking) in our case since our $\{\boldsymbol{y}(t)\}$ is not an i.i.d. sequence. □

5. THEORETICAL ANALYSIS

In this section we analyze consistency (Theorem 1) by invoking some results from [14]. The difference from [14] is that while the observations in [14] are i.i.d., here $\{\boldsymbol{x}(t)\}$, and $\{\boldsymbol{y}(t)\}$ constructed from it, are dependent sequences. Therefore, we need a model for this dependence. This influences concentration inequality regarding convergence of sample covariance $\hat{\boldsymbol{\Sigma}}$. Once this aspect is accounted for, [14, Theorem 1] applies immediately.

To quantify the dependence structure of $\{\boldsymbol{x}(t)\}$, we will follow [19]; other possibilities include [20, 21].

(A0) Assume $\{\boldsymbol{x}(t)\}$ obeys

$$\boldsymbol{x}(t) = \sum_{i=0}^{\infty} \boldsymbol{A}_i \boldsymbol{e}(t-i) \,, \tag{30}$$

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

$$\max_{1 \le q \le p} \sqrt{\sum_{k=1}^{p} ([\boldsymbol{A}_i]_{qk})^2} \le \frac{c_a}{(\max(1,i))^\gamma} \tag{31}$$

for all $i \ge 0$, some $c_a \in (0, \infty)$, and $\gamma > 1$.

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 $\|\boldsymbol{A}_i\|_F \le a|\lambda_0|^i$ for some $0 < a < \infty$ where $|\lambda_0| < 1$ is the largest magnitude "pole" (root of $c(z) := \big|\boldsymbol{I} + \sum_{i=1}^q \boldsymbol{\Phi}_i z^{-i}\big| = 0$) of the model. It can be shown that there exist $0 < b < \infty$ and $1 < \gamma < \infty$ such that $a|\lambda_0|^i \le b\, i^{-\gamma}$ for $i \ge 1$, thereby satisfying assumption (A0).

By Assumption (A0), it follows that $\boldsymbol{y}(t) = \sum_{i=0}^\infty \boldsymbol{B}_i \bar{\boldsymbol{e}}(t-i)$, $\bar{\boldsymbol{e}}(t) \in \mathbb{R}^{mp}$ 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

$$\max_{1 \le q \le mp} \sqrt{\sum_{k=1}^{mp} ([\boldsymbol{B}_i]_{qk})^2} \le \frac{c_a}{(\max(1,i))^\gamma} \tag{32}$$

for all $i \ge 0$, with c_a, and γ as in Assumption (A0). Then we have Lemma 1, following [19, Lemma VI.2, supplementary] for the case $\gamma > 1$ (γ is called β in [19]).

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

$$P\left(\max_{k,l} \Big| [\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_{y0}]_{kl} \Big| \ge \delta \right) \le 2 \, \exp(-C_u \bar{n} \min(\delta^2, \delta)) \tag{33}$$

for any $\delta > 0$, where $C_u \in (0, \infty)$ is an absolute (universal) constant. •

Constant C_u results from the application of the Hanson-Wright inequality [22].

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 leads to Lemma 2.

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

$$P\left(\max_{k,l} \Big| [\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_{y0}]_{kl} \Big| > C_0 \sqrt{\frac{\ln(mp_n)}{\bar{n}}} \right) \le \frac{1}{(mp_n)^{\tau-2}} \tag{34}$$

for $\tau > 2$, if the sample size $\bar{n} = n - d > N_1 = \ln(2(mp_n)^\tau)/C_u$, where $m = d+1$ and $C_0 = \sqrt{N_1 / \ln(mp_n)}$. •

Lemma 2 above replaces [14, Lemma 2] for dependency in observations. Further assume

(A1) Let $\boldsymbol{\Sigma}_{y0} = E\{\boldsymbol{y}(t)\boldsymbol{y}^\top(t)\} \succ \mathbf{0}$ denote the true covariance of $\boldsymbol{y}(t)$. Define $\mathcal{E}_{y0} = \{\{i,j\} \,:\, \boldsymbol{\Omega}_{y0}^{(ij)} \neq \mathbf{0}, \; i \neq j\}$ where $\boldsymbol{\Omega}_{y0} = \boldsymbol{\Sigma}_{y0}^{-1}$. Assume that $\mathrm{card}(\mathcal{E}_{y0}) = |(\mathcal{E}_0)| \le s_{n0}$.

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

$$0 < \beta_{\min} \le \phi_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{y}\mathbf{0}}) \le \phi_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{y}\mathbf{0}}) \le \beta_{\max} < \infty \,.$$

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

Let $\hat{\boldsymbol{\Omega}}_{y\lambda} = \arg\min_{\boldsymbol{\Omega}_y \succ \mathbf{0}} L_{SGL}(\boldsymbol{\Omega}_y)$. Theorem 1 establishes consistency of $\hat{\boldsymbol{\Omega}}_{y\lambda}$ and it follows by replacing [14, Lemma 2] with Lemma 2 of this paper in the proof of [14, Theorem 1].

Theorem 1 (Consistency): For $\tau > 2$, let $m = d+1$ and

$$C_0 = \sqrt{\ln(2(mp_n)^\tau)/(C_u \ln(mp_n))} \,. \tag{35}$$

Given real numbers $\delta_1 \in (0,1)$, $\delta_2 > 0$ and $C_1 > 0$, let $C_2 = \sqrt{m} + 1 + C_1$, and

$$M = (1+\delta_1)^2 (2C_2 + \delta_2) C_0 / \beta_{\min}^2, \tag{36}$$

$$r_n = \sqrt{\frac{(mp_n + m^2 s_{n0}) \ln(mp_n)}{\bar{n}}} = o(1) \,, \tag{37}$$

$$N_1 = \ln(2(mp_n)^\tau)/C_u, \tag{38}$$

$$N_2 = \arg\min \left\{ \bar{n} \,:\, r_n \le \frac{\delta_1 \beta_{\min}}{(1+\delta_1)^2 (2C_2 + \delta_2) C_0} \right\} . \tag{39}$$

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

$$\frac{C_1 C_0}{1+\alpha(m-1)}\sqrt{\Big(1+\frac{p_n}{m s_{n0}}\Big)\frac{\ln(mp_n)}{\bar{n}}} \geq \frac{\lambda_n}{m} \geq C_0\sqrt{\frac{\ln(mp_n)}{\bar{n}}}. \quad (40)$$

Then if the sample size $\bar{n} = n - d > \max\{N_1, N_2\}$ and assumptions (A0)-(A2) hold true, $\hat{\boldsymbol{\Omega}}_{y\lambda}$ satisfies

$$\|\hat{\boldsymbol{\Omega}}_{y\lambda} - \boldsymbol{\Omega}_{y0}\|_F \leq M r_n \quad (41)$$

with probability greater than $1 - 1/(mp_n)^{\tau-2}$. In terms of rate of convergence, $\|\hat{\boldsymbol{\Omega}}_{y\lambda} - \boldsymbol{\Omega}_{y0}\|_F = \mathcal{O}_P(r_n)$ •

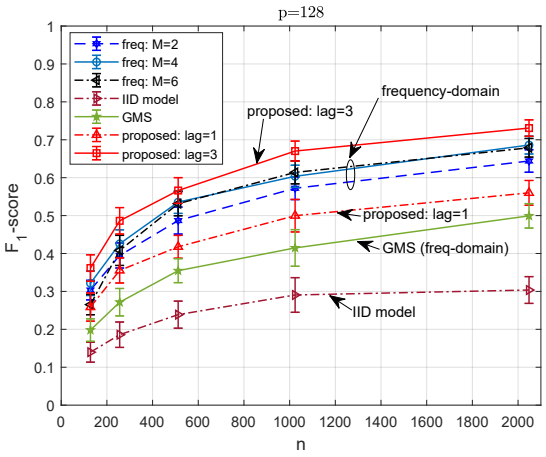

Fig. 1: F_1-scores based on 100 runs for 4 approaches. In the proposed approach lag=3 refers to $d = 3$. IID model may be viewed as proposed approach with lag=$d = 0$.

6. NUMERICAL EXAMPLE

Consider $p = 128$, 16 clusters (communities) of 8 nodes each, where nodes within a community are not connected to any nodes in other communities. Within any community of 8 nodes, the data are generated using a vector autoregressive (VAR) model of order 3. Consider community q, $q = 1, 2, \cdots, 16$. Then $\boldsymbol{x}^{(q)}(t) \in \mathbb{R}^8$ is generated as

$$\boldsymbol{x}^{(q)}(t) = \sum_{i=1}^{3} \boldsymbol{A}_i^{(q)} \boldsymbol{x}^{(q)}(t-i) + \boldsymbol{w}^{(q)}(t)$$

with $\boldsymbol{w}^{(q)}(t)$ as i.i.d. zero-mean Gaussian with identity covariance matrix. Only 10% of entries of $\boldsymbol{A}_i^{(q)}$'s are nonzero and the nonzero elements are independently and uniformly distributed over $[-0.8, 0.8]$. We then check if the VAR(3) model is stable with all eigenvalues of the companion matrix ≤ 0.95 in magnitude; if not, we re-draw randomly till this condition is fulfilled. The overall data $\boldsymbol{x}(t)$ is given by $\boldsymbol{x}(t) = [\,\boldsymbol{x}^{(1)\top}(t) \;\cdots\; \boldsymbol{x}^{(16)\top}(t)\,]^\top \in \mathbb{R}^p$. First 100 samples are discarded to eliminate transients. This set-up leads to approximately 3.5% connected edges. The true edge set $\mathcal{E}_0$ for the time series graph is determined as follows. In each run, we calculated the true PSD $\boldsymbol{S}(f)$ for $f \in [0, 0.5]$ at intervals of 0.01, and then take $\{i,j\} \in \mathcal{E}_0$ if $\sum_f |S_{ij}^{-1}(f)| > 10^{-6}$, else $\{i,j\} \notin \mathcal{E}_0$.

Simulation results based on 100 runs are shown in Figs. 1 and 2. The performance measure is F_1-score for efficacy in edge detection. The F_1-score is defined as $F_1 = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall})$ where precision $= |\hat{\mathcal{E}} \cap \mathcal{E}_0|/|\hat{\mathcal{E}}|$, recall $= |\hat{\mathcal{E}} \cap \mathcal{E}_0|/|\mathcal{E}_0|$, and $\mathcal{E}_0$ and $\hat{\mathcal{E}}$ denote the true and estimated edge sets, respectively. Four approaches were tested: **(i)** Proposed multi-attribute graph based approach with lags (delays) $d = 1$ or $d = 3$, labeled "proposed, lag=1" or "proposed, lag=3" in the figures. **(ii)** Frequency-domain sparse-group lasso approach of [11–13], optimized using ADMM, using varying number M (=2,4 or 6) of smoothed PSD estimators in frequency range (0,0.5), labeled "freq: M=2", "freq: M=4" "freq: M=6". **(iii)** An i.i.d. modeling approach that exploits only the sample covariance $\frac{1}{n}\sum_{t=0}^{n-1} \boldsymbol{x}(t)\boldsymbol{x}^\top(t)$ (labeled "IID model"), implemented via the ADMM (adaptive) lasso approach ([18, Sec. 6.4]). In this approach, as discussed in Sec. 2.1, edge $\{i,j\}$ exists in the CIG iff $\Omega_{ij} \neq 0$ where precision matrix $\boldsymbol{\Omega} = \boldsymbol{R}_{xx}^{-1}(0)$. **(iv)** The frequency-domain ADMM approach of [10], labeled "GMS" (graphical model selection), which was applied with $F = 4$ (four frequency points, corresponds to $M = 4$ in [11–13]) and all other default settings of [10] to compute the PSDs. The tuning parameters, (α, λ) for proposed and frequency-domain sparse-group lasso approach of [11–13], and lasso parameter λ for IID and GMS, were selected via an exhaustive search over a grid of values to maximize the F_1-score (which requires knowledge of the true edge-set). The results shown in Figs. 1 and 2 are based on these optimized tuning parameters. (In practice, one would use an information criterion or cross-validation to select the tuning parameters.)

The F_1-scores are shown in Fig. 1 and average timings per run are shown in Fig. 2 for sample sizes $n = 128, 256, 512, 1024, 2048$. It is seen that with F_1-score as the performance metric, our proposed method with lag $d = 3$ significantly outperforms other approaches while also being faster than frequency-domain approaches.

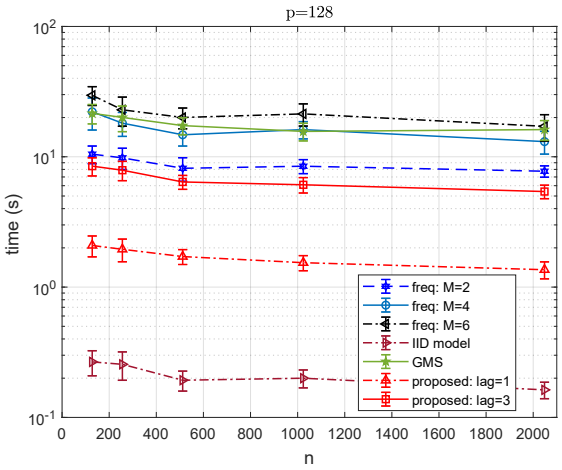

Fig. 2: Average timing per run based on 100 runs for 4 approaches.

7. CONCLUSIONS

Graphical modeling of dependent Gaussian time series was considered. We formulated the problem 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. We exploited the results of [14] to provide an ADMM solution to minimize a sparse-group lasso penalized negative pseudo log-likelihood objective function for multi-attribute graph precision matrix estimation. A theoretical analysis was provided. Numerical results were provided to illustrate the proposed approach which outperforms the approaches of [10–13] with F_1-score as the performance metric for graph edge detection.

8. 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. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol 303, pp. 799-805, 2004.

[6] S.L. Lauritzen and N.A. Sheehan, "Graphical models for genetic analyses," *Statistical Science*, vol. 18, pp. 489-514, 2003.

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

[8] 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, pp. 445-488, 2014.

[9] 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.

[10] 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.

[11] J.K. Tugnait, "Graphical modeling of high-dimensional time series," in *Proc. 52nd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Oct. 29 - Oct. 31, 2018, pp. 840-844.

[12] J.K. Tugnait, "Consistency of sparse-group lasso graphical model selection for time series," in *Proc. 54th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 1-4, 2020, pp. 589-593.

[13] J.K. Tugnait, "New results on graphical modeling of high-dimensional dependent time series," in *Proc. 55th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Oct. 31 - Nov. 3, 2021.

[14] 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.)

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

[16] J. Friedman, T. Hastie and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv:1001.0736v1 [math.ST]*, 5 Jan 2010.

[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] 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.

[19] 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. (Supplementary material available online, 18 pages.)

[20] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Annals Statistics*, vol. 43, no. 4, pp. 1535-1567, 2015.

[21] H. Shu and B. Nan, "Estimation of large covariance and precision matrices from temporally dependent observations," *Annals Statistics*, vol. 47, no. 3, pp. 1321-1350, 2019.

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