ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447097

DELAY EMBEDDING FOR MATRIX GRAPHICAL MODEL LEARNING FROM DEPENDENT DATA

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 sparse, high-dimensional, stationary matrix-variate Gaussian time series. The correlation function of the matrix series is Kronecker-decomposable. Unlike most past work on matrix graphical models, where independent and identically distributed (i.i.d.) observations of matrix-variate are assumed to be available, we allow time-dependent observations. We follow a time-delay embedding approach where with each matrix node, we associate a random vector consisting of a scalar series component and its time-delayed copies. A group-lasso penalized negative pseudo log-likelihood (NPLL) objective function is formulated to estimate a Kronecker-decomposable covariance matrix which allows for inference of the underlying CIG. The NPLL function is bi-convex and the Kronecker-decomposable covariance matrix is estimated via flip-flop optimization of the NPLL function. Each iteration of flip-flop optimization is solved via an alternating direction method of multipliers (ADMM) approach. Numerical results illustrate the proposed approach which outperforms an existing i.i.d. modeling based approach as well as an existing frequency-domain approach for dependent data, in correctly detecting the graph edges.

Keywords: Sparse graph learning; matrix graph estimation; matrix time series; undirected graph; delay embedding.

1. INTRODUCTION

In graphical models, graphs display the conditional independence structure of the random variables [1]. While vector graphical models have been extensively studied [2–4], much less attention has been given to matrix-valued graphical models, the need for which arises in several applications [5–14].

In a vector graphical model, the conditional statistical dependency structure among p random variables $x_1, x_1, \cdots, x_p$, is represented using an undirected 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 $\mathcal{E} \subseteq [p] \times [p]$. There is no edge between nodes i and j iff x_i and x_j are conditionally independent given the remaining p-2 variables. Suppose $\boldsymbol{x}$ has positive-definite covariance matrix $\boldsymbol{\Sigma}$ with precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Then Ω_{ij}, the (i,j)-th element of $\boldsymbol{\Omega}$, is zero iff x_i and x_j are conditionally independent [1]. In matrix graphs, we observe matrix-valued time series $\{\boldsymbol{Z}(t)\}$ where $\boldsymbol{Z}(t) \in \mathbb{R}^{p\times q}$. If one vectorizes using vec($\boldsymbol{Z}$), then use of vec($\boldsymbol{Z}$) will result in a pq-node vector graph with $(pq) \times (pq)$ precision matrix, which could be ultra-high-dimensional and moreover, it ignores any structural information among rows and columns of the matrix observations [5]. The basic idea in matrix-valued graphs is to model the covariance of vec($\boldsymbol{Z}$) as $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$ ($\boldsymbol{\Psi}$ is $q \times q$ and $\boldsymbol{\Sigma}$ is $p \times p$), reducing the number of unknowns from $\mathcal{O}(p^2q^2)$ in the precision matrix for the "full" vectorized model to $\mathcal{O}(p^2 + q^2)$ for the matrix model, while also preserving the structural information.

Prior work [5–14] on matrix (or tensor) graphs all assume that i.i.d. observations of $\boldsymbol{Z}$ are available for graphical modeling. Our objective in this paper is to learn the matrix graph associated with time-dependent matrix-valued $p \times q$ Gaussian sequence $\boldsymbol{Z}(t)$, given observations of $\boldsymbol{Z}(t)$ for $t = 1, 2, \cdots, n$.

Notation: $|\boldsymbol{A}|$ and $\mathrm{tr}(\boldsymbol{A})$ denote the determinant and the trace of the square matrix $\boldsymbol{A}$, respectively, and $\mathrm{etr}(\boldsymbol{A}) = \exp(\mathrm{tr}(\boldsymbol{A}))$. $[\boldsymbol{B}]_{ij}$ denotes the (i,j)-th element of $\boldsymbol{B}$, and so does B_{ij}. $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$ denotes a Gaussian random vector $\boldsymbol{x}$ with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$, and $\otimes$ denotes the Kronecker product.

2. SYSTEM MODEL AND PRIOR WORK

Random matrix $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$ is said to have a matrix normal (Gaussian) distribution if its pdf $f(\boldsymbol{Z}|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$, characterized by $\boldsymbol{M} \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$, is given by [15, Chap. 2]

$$f(\boldsymbol{Z}|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = \frac{\mathrm{etr}\Big(-\frac{1}{2}(\boldsymbol{Z}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Z}-\boldsymbol{M})^\top\boldsymbol{\Sigma}^{-1}\Big)}{(2\pi)^{pq/2}|\boldsymbol{\Sigma}|^{q/2}|\boldsymbol{\Psi}|^{p/2}}, \tag{1}$$

where $\mathrm{etr}(\boldsymbol{A}) = \exp(\mathrm{tr}(\boldsymbol{A}))$. Equivalently, $\boldsymbol{z} = \mathrm{vec}(\boldsymbol{Z}) \sim \mathcal{N}\big(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}\big)$. Here $\boldsymbol{\Psi}$ is the row covariance matrix and $\boldsymbol{\Sigma}$ is the column covariance matrix [15] since the kth column $\boldsymbol{Z}_{\cdot k} \sim \mathcal{N}(\boldsymbol{0}, [\boldsymbol{\Psi}]_{kk}\boldsymbol{\Sigma})$ and the ith row $\boldsymbol{Z}_{i\cdot}^\top \sim \mathcal{N}(\boldsymbol{0}, [\boldsymbol{\Sigma}]_{ii}\boldsymbol{\Psi})$.

Graphical modeling of random vectors to characterize conditional dependence of its components [1, 2] was extended to matrix data with structured information [5, 7, 8, 11, 12]. With $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$ modeled as zero-mean matrix normal, and $\boldsymbol{z} = \mathrm{vec}(\boldsymbol{Z})$, we have $E\{\boldsymbol{z}\boldsymbol{z}^\top\} = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$, implying a separable covariance structure [16]. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Gamma} = \boldsymbol{\Psi}^{-1}$ denote the respective precision matrices. Then $\boldsymbol{Z}_{ij}$ and $\boldsymbol{Z}_{k\ell}$ are conditionally independent given remaining entries in $\boldsymbol{Z}$ iff (i) at least one of $\boldsymbol{\Omega}_{ik}$ and $\boldsymbol{\Gamma}_{j\ell}$ is zero when $i \neq k$, $j \neq \ell$, (ii) $\boldsymbol{\Omega}_{ik} = 0$ when $i \neq k$, $j = \ell$, and (iii) $\boldsymbol{\Gamma}_{j\ell} = 0$ when $i = k$, $j \neq \ell$ [5]. Prior work [5–14] assumes that i.i.d. observations of $\boldsymbol{Z}$ are available for graphical modeling. Recently in [17] time-dependence was introduced, modeling the time-dependent zero-mean matrix-valued, stationary, $p \times q$ Gaussian sequence $\boldsymbol{Z}(t)$, $\boldsymbol{z}(t) = \mathrm{vec}(\boldsymbol{Z}(t))$, as having the separable covariance structure given by

$$E\{\boldsymbol{z}(t+\tau)\boldsymbol{z}^\top(t)\} = \boldsymbol{\Psi}(\tau) \otimes \boldsymbol{\Sigma} \tag{2}$$

where $\boldsymbol{\Psi}(\tau)$, $\tau = 0, \pm 1, \cdots$ models time-dependence while $\boldsymbol{\Sigma} \succ \boldsymbol{0}$ is fixed. With $\{\boldsymbol{e}(t)\}$ i.i.d., $\boldsymbol{e}(t) \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{I}_{pq})$, a generative model

This work was supported by NSF Grant CCF-2308473. Author's email: tugnajk@auburn.edu

for $\boldsymbol{z}(t)$ is given by

$$\boldsymbol{z}(t) = \sum_{i=0}^{L} (\boldsymbol{B}_i \otimes \boldsymbol{F}) \boldsymbol{e}(t-i)\,,\ \boldsymbol{B}_i \in \mathbb{R}^{q\times q}\,,\ \boldsymbol{F} \in \mathbb{R}^{p\times p}\,, \quad (3)$$

$$\boldsymbol{\Psi}(\tau) = \sum_{i=0}^{L} \boldsymbol{B}_i \boldsymbol{B}_{i-\tau}^{\top} \quad \text{and} \quad \boldsymbol{\Sigma} = \boldsymbol{F}\boldsymbol{F}^{\top}\,. \quad (4)$$

The power spectral density (PSD) of $\{\boldsymbol{z}(t)\}$ is $\boldsymbol{S}_z(f) = \bar{\boldsymbol{S}}(f) \otimes \boldsymbol{\Sigma}$ where $\bar{\boldsymbol{S}}(f) = \sum_\tau \boldsymbol{\Psi}(\tau) e^{-j2\pi f\tau}$. Then $\boldsymbol{S}_z^{-1}(f) = \bar{\boldsymbol{S}}^{-1}(f) \otimes \boldsymbol{\Sigma}^{-1}$, and by [18], in the $pq-$node graph $\mathcal{G} = (V, \mathcal{E})$, $|V| = pq$, associated with $\{\boldsymbol{z}(t)\}$, edge $\{i,j\} \notin \mathcal{E}$ iff $[\boldsymbol{S}_z^{-1}(f)]_{ij} = 0$ for every f. This does not account for the separable structure of the model. Noting that $\bar{\boldsymbol{S}}^{-1}(f)$, $f \in [0, 0.5]$, plays the role of $\Gamma = \boldsymbol{\Psi}^{-1}$, using [5, 18], it follows that $\{\boldsymbol{Z}_{ij}(t)\}$ and $\{\boldsymbol{Z}_{k\ell}(t)\}$ are conditionally independent given remaining entries in $\{\boldsymbol{Z}(t)\}$ iff (i) at least one of $\boldsymbol{\Omega}_{ik}$ and $[\bar{\boldsymbol{S}}^{-1}(f)]_{j\ell}$, $f \in [0, 0.5]$ is zero when $i \neq k$, $j \neq \ell$, (ii) $\boldsymbol{\Omega}_{ik} = 0$ when $i \neq k$, $j = \ell$, and (iii) $[\bar{\boldsymbol{S}}^{-1}(f)]_{j\ell} = 0$ for $f \in [0, 0.5]$ when $i = k$, $j \neq \ell$.

In [17] the objective was to learn the graph associated with time-dependent sequence $\{\boldsymbol{Z}(t)\}$, given observations $t = 1, 2, \cdots, n$, under some sparsity constraints on $\boldsymbol{\Omega}$ and $\bar{\boldsymbol{S}}^{-1}(f)$, $f \in [0, 0.5]$. In this paper we follow a time-domain approach using a time-delay embedding approach (called a "multi-attribute formulation" in [19] in the context of graphical modeling of dependent time series). The name delay embedding originates from [20] in the context of chaos detection where such embeddings reveal the dynamic structure of the underlying system, and such embeddings have been employed in other contexts, e.g., [21, 22].

2.1. Gaussian Graphical Models

2.1.1. Vector Graphical Model

Given an undirected graph $\mathcal{G} = (V, \mathcal{E})$, $|V| = p$, in a vector graphical model for zero-mean Gaussian $\boldsymbol{x} \in \mathbb{R}^p$, component x_i is associated with node i of the graph, and the conditional independence relationships among x_i's are encoded in $\mathcal{E}$. Let $\boldsymbol{x}_{-ij} = \{x_k\,:\, k \in V\backslash\{i,j\}\} \in \mathbb{R}^{p-2}$ denote the vector $\boldsymbol{x}$ after deleting x_i and x_j from it where $V\backslash\{i,j\}$ denotes the set V with nodes i and j deleted. Define $e_{i|-ij} = x_i - E\{x_i|\boldsymbol{x}_{-ij}\}$ and $e_{j|-ij} = x_j - E\{x_j|\boldsymbol{x}_{-ij}\}$. Then we have the following equivalence [1]

$$\text{edge } \{i,j\} \notin \mathcal{E} \ \Leftrightarrow\ \Omega_{ij} = 0 \ \Leftrightarrow\ E\{e_{i|-ij} e_{j|-ij}\} = 0\,. \quad (5)$$

2.1.2. Vector Time Series Graphical Model

In this model for a stationary zero-mean Gaussian time series $\{\boldsymbol{x}(t)\}$, $\boldsymbol{x}(t) \in \mathbb{R}^p$, component series $\{x_i(t)\}$ is associated with node i of the graph, and the conditional independence relationships among series components are encoded in $\mathcal{E}$. Define $e_{i|-ij}(t) = x_i(t) - E\{x_i(t)|\boldsymbol{x}_{-ij,\mathbb{Z}}\}$, $e_{j|-ij}(t) = x_j(t) - E\{x_j(t)|\boldsymbol{x}_{-ij,\mathbb{Z}}\}$ where $\boldsymbol{x}_{-ij,\mathbb{Z}} = \{x_k(t)\,:\, k \in V\backslash\{i,j\},\ t \in \mathbb{Z}\}$. Then we have the following equivalence [18] ($\mathbb{Z}$ is the set of integers)

$$\begin{aligned}\text{edge } \{i,j\} \notin \mathcal{E} \ &\Leftrightarrow\ [\boldsymbol{S}_x^{-1}(f)]_{ij} = 0\ \forall f \in [0,1] \\ &\Leftrightarrow\ E\{e_{i|-ij}(t+\tau) e_{j|-ij}(t)\} = 0\ \forall \tau \in \mathbb{Z}\,. \end{aligned} \quad (6)$$

2.1.3. Vector Multi-Attribute Graphical Model

In this model for p jointly Gaussian vectors $\boldsymbol{z}_i \in \mathbb{R}^m$, $i \in [p]$, $\boldsymbol{z}_i$ is associated with node i of $\mathcal{G} = (V, \mathcal{E})$, $V = [p]$. We now have m attributes per node. Let $\boldsymbol{x} = [\boldsymbol{z}_1^{\top}\ \boldsymbol{z}_2^{\top}\ \cdots\ \boldsymbol{z}_p^{\top}]^{\top} \in \mathbb{R}^{mp}$. 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

$$[\boldsymbol{\Omega}^{(ij)}]_{rs} = [\boldsymbol{\Omega}]_{(i-1)m+r,(j-1)m+s}\,,\ r, s = 1, 2, \cdots, m\,. \quad (7)$$

Let $\boldsymbol{z}_{-ij} = \{\boldsymbol{z}_k\,:\, k \in V\backslash\{i,j\}\} \in \mathbb{R}^{m(p-2)}$ denote the vector $\boldsymbol{x}$ after deleting vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ from it. Define $\boldsymbol{e}_{i|-ij} = \boldsymbol{z}_i - E\{\boldsymbol{z}_i|\boldsymbol{z}_{-ij}\}$ and $\boldsymbol{e}_{j|-ij} = \boldsymbol{z}_j - E\{\boldsymbol{z}_j|\boldsymbol{z}_{-ij}\}$. Then we have the following equivalence [23, Sec. 2.1, Appendix B.3]

$$\text{edge } \{i,j\} \notin \mathcal{E} \ \Leftrightarrow\ \boldsymbol{\Omega}^{(ij)} = \boldsymbol{0} \ \Leftrightarrow\ E\{\boldsymbol{e}_{i|-ij}\boldsymbol{e}_{j|-ij}^{\top}\} = \boldsymbol{0}\,. \quad (8)$$

3. TIME-DELAY EMBEDDING

Define the time-delay embedded vector $\boldsymbol{y}(t) \in \mathbb{R}^{mpq}$ as

$$\boldsymbol{y}(t) = [\boldsymbol{z}^{\top}(t),\ \boldsymbol{z}^{\top}(t-1),\ \cdots,\ \boldsymbol{z}^{\top}(t-(m-1))]^{\top} \quad (9)$$

and the corresponding delay embedded matrix $\boldsymbol{Y}(t) \in \mathbb{R}^{p\times mq}$

$$\boldsymbol{Y}(t) = [\boldsymbol{Z}(t),\ \boldsymbol{Z}(t-1),\ \cdots,\ \boldsymbol{Z}(t-(m-1))] \quad (10)$$

such that $\boldsymbol{y}(t) = \text{vec}(\boldsymbol{Y}(t))$. By assumption, $\{\boldsymbol{Z}(t)\}$ is a matrix normal sequence with Kronecker-decomposable covariance structure specified by (2). We have

$$E\{\boldsymbol{y}(t)\boldsymbol{y}^{\top}(t)\} = \tilde{\boldsymbol{\Psi}} \otimes \boldsymbol{\Sigma}\,,\ \tilde{\boldsymbol{\Psi}} \in \mathbb{R}^{mq\times mq}\,,\ \tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{p\times p}\,, \quad (11)$$

$$\tilde{\boldsymbol{\Psi}} = \begin{bmatrix} \boldsymbol{\Psi}(0) & \boldsymbol{\Psi}(1) & \cdots & \boldsymbol{\Psi}(m-1) \\ \boldsymbol{\Psi}(-1) & \boldsymbol{\Psi}(0) & \cdots & \boldsymbol{\Psi}(m-2) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Psi}(-m+1) & \boldsymbol{\Psi}(-m+2) & \cdots & \boldsymbol{\Psi}(0) \end{bmatrix} \quad (12)$$

Define $\tilde{\boldsymbol{\Gamma}} = \tilde{\boldsymbol{\Psi}}^{-1}$, and let, as before, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Define the $m \times m$ matrix $\tilde{\boldsymbol{\Gamma}}^{(k\ell)}$, comprised of certain m^2 elements of $\tilde{\boldsymbol{\Gamma}}$, with (r,s)th element of $\tilde{\boldsymbol{\Gamma}}^{(k\ell)}$ as

$$[\tilde{\boldsymbol{\Gamma}}^{(k\ell)}]_{rs} = [\tilde{\boldsymbol{\Gamma}}]_{k+(r-1)q,\ell+(s-1)q}, \quad r, s \in [m]\,. \quad (13)$$

Lemma 1. Suppose that the stationary zero-mean matrix normal sequence $\{\boldsymbol{Z}(t)\}$ is generated via (3). Then $\{\boldsymbol{Z}_{ij}(u),\ t-m+1 \le u \le t\}$ and $\{\boldsymbol{Z}_{k\ell}(u),\ t-m+1 \le u \le t\}$ are conditionally independent given $\{\boldsymbol{Z}_{rs}(u),\ t-m+1 \le u \le t,\ (r,s) \notin \{(i,j),(k,\ell)\},\ r \in [p],\ s \in [q]\}$ iff

(i) at least one of $\boldsymbol{\Omega}_{ik}$ and $\|\tilde{\boldsymbol{\Gamma}}^{(j\ell)}\|_F$ is zero when $i \neq k$, $j \neq \ell$,

(ii) $\boldsymbol{\Omega}_{ik} = 0$ when $i \neq k$, $j = \ell$,

(iii) $\|\tilde{\boldsymbol{\Gamma}}^{(j\ell)}\|_F = 0$ when $i = k$, $j \neq \ell$. •

Proof. Associate $\boldsymbol{z}(t) = \text{vec}(\boldsymbol{Z}(t))$ with graph $\mathcal{G} = (V, \mathcal{E})$, $|V| = pq$. In order to exploit the formulation of Sec. 2.1.3, define

$$\tilde{\boldsymbol{z}}^{(i)}(t) = [\boldsymbol{z}_i(t),\ \boldsymbol{z}_i(t-1),\ \cdots,\ \boldsymbol{z}_i(t-(m-1))]^{\top}\,, \quad (14)$$

$$\tilde{\boldsymbol{y}}(t) = [(\tilde{\boldsymbol{z}}^{(1)}(t))^{\top},\ (\tilde{\boldsymbol{z}}^{(2)}(t))^{\top},\ \cdots,\ (\tilde{\boldsymbol{z}}^{(pq)}(t))^{\top}]^{\top} \quad (15)$$

where $\tilde{\boldsymbol{z}}^{(i)}(t) \in \mathbb{R}^m$ and $\tilde{\boldsymbol{y}}(t) \in \mathbb{R}^{mpq}$. The graph $\mathcal{G} = (V, \mathcal{E})$ also describes $\tilde{\boldsymbol{y}}(t)$ as a multi-attribute graphical model. Original $\boldsymbol{y}(t)$ and new $\tilde{\boldsymbol{y}}(t)$ are related by an $mpq \times mpq$ permutation matrix $\boldsymbol{P}$ with $\tilde{\boldsymbol{y}}(t) = \boldsymbol{P}\boldsymbol{y}(t)$. Let

$$\tilde{\boldsymbol{z}}_{-vw}(t) = \{\tilde{\boldsymbol{z}}^{(a)}(t)\,:\, a \in \tilde{V}\backslash\{v,w\}\}\,,\ \ v, w \in [pq]\,, \quad (16)$$

$$\boldsymbol{e}_{v|-vw}(t) = \tilde{\boldsymbol{z}}^{(v)}(t) - E\{\tilde{\boldsymbol{z}}^{(v)}(t)|\tilde{\boldsymbol{z}}_{-vw}(t)\}\,, \quad (17)$$

$$\boldsymbol{e}_{w|-vw}(t) = \tilde{\boldsymbol{z}}^{(w)}(t) - E\{\tilde{\boldsymbol{z}}^{(w}(t)|\tilde{\boldsymbol{z}}_{-vw}(t)\}\,. \quad (18)$$

By Sec. 2.1.3 ($\tilde{\boldsymbol{\Omega}}_y^{(vw)}$ is similar to (7) except that now $|V| = pq$),

$$\{v,w\} \notin \mathcal{E} \;\Leftrightarrow\; \tilde{\boldsymbol{\Omega}}_y^{(vw)} = \mathbf{0} \tag{19}$$

where $\tilde{\boldsymbol{\Omega}}_y = (E\{\tilde{\boldsymbol{y}}(t)\tilde{\boldsymbol{y}}^\top(t)\})^{-1}$. Define

$$\check{\boldsymbol{z}}_{-vw;t,m} = \{z_a(s) \,:\, a \in \tilde{V}\backslash\{v,w\},\; t-m+1 \le s \le t\}\,, \tag{20}$$

$$e_{zv|-vw}(t') = z_v(t') - E\{z_v(t')|\check{\boldsymbol{z}}_{-vw;t,m}\}\,, \tag{21}$$

$$e_{zw|-vw}(t') = z_v(t') - E\{z_w(t')|\check{\boldsymbol{z}}_{-vw;t,m}\}\,. \tag{22}$$

Notice that $e_{zv|-vw}(t')$ above is an element of $\boldsymbol{e}_{v|-vw}(t)$ defined in (17) for any $t-m+1 \le t' \le t$. Then by (8) and (19), we have

$$\tilde{\boldsymbol{\Omega}}_y^{(vw)} = \mathbf{0} \;\Leftrightarrow\; E\{e_{zv|-vw}(t_1)e_{zw|-vw}(t_2)\} = 0, \quad \text{for} \;\; t-m+1 \le t_1, t_2 \le t. \tag{23}$$

With $\boldsymbol{\Omega}_y = (E\{\boldsymbol{y}(t)\boldsymbol{y}^\top(t)\})^{-1} = (\tilde{\boldsymbol{\Psi}} \otimes \boldsymbol{\Sigma})^{-1} = \tilde{\boldsymbol{\Gamma}} \otimes \boldsymbol{\Omega}$, we have $\tilde{\boldsymbol{\Omega}}_y = \boldsymbol{P}(\tilde{\boldsymbol{\Gamma}} \otimes \boldsymbol{\Omega})\boldsymbol{P}^\top$ since $\tilde{\boldsymbol{y}}(t) = \boldsymbol{P}\boldsymbol{y}(t)$ and $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$. Hence $\tilde{\boldsymbol{\Omega}}_y^{(vw)} = \mathbf{0} \;\Leftrightarrow\; (\boldsymbol{P}(\tilde{\boldsymbol{\Gamma}} \otimes \boldsymbol{\Omega})\boldsymbol{P}^\top)^{(vw)} = \mathbf{0} \;\Leftrightarrow\; \{v,w\} \notin \mathcal{E}$. The conclusions of Lemma 1 parts (i)-(iii) then follow by using the Kronecker product structure $\tilde{\boldsymbol{\Gamma}} \otimes \boldsymbol{\Omega}$, and exploiting the correspondence between the entries of $\tilde{\boldsymbol{y}}(t)$ and $\boldsymbol{y}(t)$. □

Remark 1. If we let $m \uparrow \infty$, the Lemma 1 implies that checking if $\|\tilde{\boldsymbol{\Gamma}}^{(k\ell)}\|_F = 0$ and/or $\boldsymbol{\Omega}_{ij} = 0$ to ascertain (19) becomes a surrogate for checking if the last equivalence in (6) holds true for graph structure estimation for time series $\{\text{vec}(\boldsymbol{Z}(t)\}$ without using frequency-domain methods. In (23), $|\tau| = |t_1 - t_2| \le m-1$, and as $m \uparrow \infty$, we approach (6) for edge $\{u,w\}$ of the graph for $\{\text{vec}(\boldsymbol{Z}(t)\}$. □

4. PENALIZED PSEUDO LOG-LIKELIHOOD

Given data $\boldsymbol{Z}(t)$, $t = 1, 2, \cdots, n$, form $\boldsymbol{Y}(t)$ as in (10) for $t = m, m+1, \cdots, n$. By (11), $\boldsymbol{y}(t) = \text{vec}(\boldsymbol{Y}(t)) \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Psi}} \otimes \boldsymbol{\Sigma})$. Therefore, pdf of $\boldsymbol{Y}(t)$ is given by

$$f_{\boldsymbol{Y}(t)}(\boldsymbol{Y}(t)) = \frac{\text{etr}\Big(-\frac{1}{2}\boldsymbol{Y}(t)\tilde{\boldsymbol{\Gamma}}\boldsymbol{Y}^\top(t)\boldsymbol{\Omega}\Big)}{(2\pi)^{mpq/2}|\boldsymbol{\Sigma}|^{mq/2}|\tilde{\boldsymbol{\Psi}}|^{p/2}}\,. \tag{24}$$

However, $\{\boldsymbol{Y}(t)\}$ is not an independent sequence. We will pretend that is an i.i.d. sequence and define a pseudo likelihood function for dataset $\mathcal{Y} = \{\boldsymbol{Y}(t)\}_{t=m}^n$ as $f_{\mathcal{Y}}(\mathcal{Y}) = \prod_{t=m}^n f_{\boldsymbol{Y}(t)}(\boldsymbol{Y}(t))$, resulting in a negative pseudo log-likelihood (NPLL) function $L(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Gamma}}) \propto -\ln(f_{\mathcal{Y}}(\mathcal{Y}))$, up to a constant, as ($n_s = n-m+1$)

$$L(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Gamma}}) = -\frac{mqn_s}{2}\ln(|\boldsymbol{\Omega}|) - \frac{pn_s}{2}\ln(|\tilde{\boldsymbol{\Gamma}}|) + \frac{1}{2}\sum_{t=m}^n \text{tr}\Big(\boldsymbol{Y}(t)\tilde{\boldsymbol{\Gamma}}\boldsymbol{Y}^\top(t)\boldsymbol{\Omega}\Big)\,. \tag{25}$$

In the high-dimension case, one needs to use penalty terms to enforce sparsity and to make the problem well-conditioned. Our proposed penalized (scaled) NPLL function is

$$\mathcal{L}(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Gamma}}) = \frac{1}{n_s mqp}\sum_{t=m}^n \text{tr}\Big(\boldsymbol{Y}(t)\tilde{\boldsymbol{\Gamma}}\boldsymbol{Y}^\top(t)\boldsymbol{\Omega}\Big) - \frac{1}{p}\ln(|\boldsymbol{\Omega}|) - \frac{1}{mq}\ln(|\tilde{\boldsymbol{\Gamma}}|) + \lambda_p \sum_{i,j=1}^p |\boldsymbol{\Omega}_{ij}| + \sqrt{m}\,\lambda_q \sum_{k,\ell=1}^q \|\tilde{\boldsymbol{\Gamma}}^{(k\ell)}\|_F \tag{26}$$

where $\lambda_p, \lambda_q > 0$ are tuning parameters, we have lasso penalty on $\boldsymbol{\Omega}$ and group lasso penalty on $\tilde{\boldsymbol{\Gamma}}$ with $\sqrt{m}$ reflecting the number of group variables. The cost (26) modifies the cost in [5] to allow for delay embeddings resulting in group lasso.

5. OPTIMIZATION

The objective function $\mathcal{L}(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Gamma}})$ in (26) is biconvex: (strictly) convex in $\tilde{\boldsymbol{\Gamma}}$, $\tilde{\boldsymbol{\Gamma}} \succ \mathbf{0}$, for fixed $\boldsymbol{\Omega}$, and (strictly) convex in $\boldsymbol{\Omega}$, $\boldsymbol{\Omega} \succ \mathbf{0}$, for fixed $\tilde{\boldsymbol{\Gamma}}$. As in [5, 7] (and others) pertaining to the i.i.d. observations case, and as is a general approach for biconvex function optimization [24, Sec. 4.2.1], we will use an iterative and alternating minimization approach where we optimize w.r.t. $\boldsymbol{\Omega}$ with $\tilde{\boldsymbol{\Gamma}}$ fixed, and then optimize w.r.t. $\tilde{\boldsymbol{\Gamma}}$ with $\boldsymbol{\Omega}$ fixed at the last optimized value, and repeat the two optimizations (flip-flop). There is no guarantee that the algorithm converges to the global minimum, however, the algorithm converges to a local stationary point of $\mathcal{L}(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Gamma}})$ [24, Sec. 4.2.1].

With $\hat{\tilde{\boldsymbol{\Gamma}}}$ denoting the estimate of $\tilde{\boldsymbol{\Gamma}}$, fix $\tilde{\boldsymbol{\Gamma}} = \hat{\tilde{\boldsymbol{\Gamma}}}$ and let $\mathcal{L}_1(\boldsymbol{\Omega})$ denote $\mathcal{L}(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Gamma}})$ up to some irrelevant constants. We minimize $\mathcal{L}_1(\boldsymbol{\Omega})$ w.r.t. $\boldsymbol{\Omega}$ to obtain estimate $\hat{\boldsymbol{\Omega}}$, where

$$\mathcal{L}_1(\boldsymbol{\Omega}) = -\frac{1}{p}\ln(|\boldsymbol{\Omega}|) + \frac{1}{p}\text{tr}\left(\boldsymbol{\Omega}\bar{\boldsymbol{S}}\right) + \lambda_p \sum_{i,j=1}^p |\boldsymbol{\Omega}_{ij}|\,, \tag{27}$$

$$\bar{\boldsymbol{S}} = \frac{1}{n_s mq}\sum_{t=m}^n \boldsymbol{Y}(t)\hat{\tilde{\boldsymbol{\Gamma}}}\boldsymbol{Y}^\top(t)\,, \quad n_s = n-m+1\,. \tag{28}$$

Fix $\boldsymbol{\Omega} = \hat{\boldsymbol{\Omega}}$ and and let $\mathcal{L}_2(\tilde{\boldsymbol{\Gamma}})$ denote $\mathcal{L}(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Gamma}})$ up to some irrelevant constants. We minimize $\mathcal{L}_2(\tilde{\boldsymbol{\Gamma}})$ w.r.t. $\tilde{\boldsymbol{\Gamma}}$ to obtain estimate $\hat{\tilde{\boldsymbol{\Gamma}}}$, where

$$\mathcal{L}_2(\tilde{\boldsymbol{\Gamma}}) = -\frac{1}{mq}\ln(|\tilde{\boldsymbol{\Gamma}}|) + \frac{1}{mq}\text{tr}\left(\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{S}}\right) + \sqrt{m}\,\lambda_q \sum_{k,\ell=1}^q \|\tilde{\boldsymbol{\Gamma}}^{(k\ell)}\|_F\,, \tag{29}$$

$$\tilde{\boldsymbol{S}} = \frac{1}{n_s p}\sum_{t=m}^n \boldsymbol{Y}^\top(t)\hat{\boldsymbol{\Omega}}\boldsymbol{Y}(t)\,. \tag{30}$$

Our optimization algorithm (used in our simulations) is as follows for a pre-chosen $m > 1$ (maximum time delay $m-1$).

1. Initialize $r = 1$, $\boldsymbol{\Omega}^{(0)} = \boldsymbol{I}_p$, $\tilde{\boldsymbol{\Gamma}}^{(0)} = \boldsymbol{I}_{mq}$.
2. Set $\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}^{(r-1)}$ in (30). Use the iterative alternating direction method of multipliers (ADMM) algorithm [25] to minimize $\mathcal{L}_2(\tilde{\boldsymbol{\Gamma}})$ (given by (29)) w.r.t. $\tilde{\boldsymbol{\Gamma}}$ to obtain estimates $\tilde{\boldsymbol{\Gamma}}^{(r)}$. [We used the ADMM algorithm of [26, Sec. III] (with $\alpha = 0$ therein, no lasso penalty). Cost (7) in [26] (after setting $\alpha = 0$) corresponds to (28) of this paper.]
3. Set $\tilde{\boldsymbol{\Gamma}} = \tilde{\boldsymbol{\Gamma}}^{(r)}$ in (28). Use the ADMM algorithm to minimize $\mathcal{L}_1(\boldsymbol{\Omega})$ w.r.t. $\boldsymbol{\Omega}$, to obtain estimate $\boldsymbol{\Omega}^{(r)}$. [We used the ADMM algorithm of [26, Sec. III] (with $\alpha = 1$ therein, no group-lasso penalty). Cost (7) in [26] (after setting $\alpha = 1$) corresponds to (27) of this paper.]
4. To resolve a scaling ambiguity, set $\boldsymbol{\Omega}^{(r)} = \boldsymbol{\Omega}^{(r)}/\|\boldsymbol{\Omega}^{(r)}\|_F$.
5. Repeat steps 2 and 4 until convergence.

6. NUMERICAL RESULTS

We use model (3)-(4) to generate synthetic data where $\boldsymbol{\Psi}(\tau)$ is controlled via a vector autoregressive (VAR) model impulse response and $\boldsymbol{\Sigma}$ is determined via an Erdös-Rènyi graph. We take $p = q = 15$. Consider the impulse response $\boldsymbol{H}^{(r)}(t) \in \mathbb{R}^{5\times 5}$ generated as $\boldsymbol{H}_i^{(r)} = \sum_{k=1}^3 \boldsymbol{A}_k^{(r)}\boldsymbol{H}_{i-k}^{(r)} + \boldsymbol{I}_5\delta_i$, where $\boldsymbol{H}_i^{(r)} = 0$ for $i < 0$, δ_i is the Kronecker delta, $r = 1, 2, 3$, and only 5% of entries of $\boldsymbol{A}_i^{(r)}$'s are nonzero and the nonzero elements are independently and

Approach	F_1 score ($\pm\sigma$)	timing (s) ($\pm\sigma$)	TPR ($\pm\sigma$)	1-TNR ($\pm\sigma$)
		$n = 64$		
IID [5–7]	0.3970 $\pm$0.1119	0.0052 $\pm$0.0011	0.2842 $\pm$0.1031	0.0015 $\pm$0.0014
Freq-domain [17]	0.7320 $\pm$0.1056	0.2001 $\pm$0.0393	0.6321 $\pm$0.1438	0.0009 $\pm$0.0011
Delay Embedding (proposed)	0.8122 $\pm$0.0792	0.0635 $\pm$0.0189	0.7450 $\pm$0.1291	0.0010 $\pm$0.0013
		$n = 256$		
IID [5–7]	0.4383 $\pm$0.1323	0.0122 $\pm$0.0042	0.3068 $\pm$0.1191	0.0008 $\pm$0.0008
Freq-domain [17]	0.8154 $\pm$0.0911	0.2278 $\pm$0.0340	0.7782 $\pm$0.1323	0.0017 $\pm$0.0015
Delay Embedding (proposed)	0.8722 $\pm$0.0767	0.1181 $\pm$0.0623	0.8693 $\pm$0.1016	0.0017 $\pm$0.0017

Table 1: *Comparisons among three approaches:* $n = 64,\ 256$, $p = q = 15$. *Tuning parameters* λ_p, λ_q *picked to yield the highest* F_1 *score. Results based on 100 runs.*

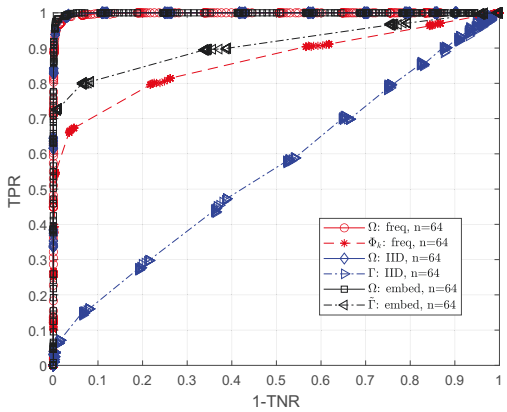

Fig. 1: ROC curves for $n = 64$: plots labeled "IID" are from the approach of [5–7], those labeled "freq." are from [17], and the ones labeled "embed" are from our proposed approach. TPR=true positive rate, TNR=true negative rate

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 impulse response $\boldsymbol{B}_i \in \mathbb{R}^{15\times 15}$ in (3) is given by $\boldsymbol{B}_i = \text{block-diag}\{\boldsymbol{H}_i^{(1)}, \boldsymbol{H}_i^{(2)}, \boldsymbol{H}_i^{(3)}\}$, for $0 \leq i \leq L = 40$, otherwise it is set to zero. Thus $\boldsymbol{B}_i$'s in (3) have a block-diagonal structure with 3 blocks, each block is 5×5. In the Erdös-Rènyi graph with $p = 15$ nodes, the nodes are connected with probability $p_{er} = 0.05$. In the upper triangular $\bar{\boldsymbol{\Omega}}$, $\bar{\boldsymbol{\Omega}}_{ij} = 0$ if $\{i, j\} \notin \mathcal{S}_p$, $\bar{\boldsymbol{\Omega}}_{ij}$ is uniformly distributed over $[-0.4, -0.1] \cup [0.1, 0.4]$ if $\{i, j\} \in \mathcal{S}_p$, and $\bar{\boldsymbol{\Omega}}_{ii} = 0.5$. With $\bar{\boldsymbol{\Omega}} = \bar{\boldsymbol{\Omega}}^\top$, add $\kappa \boldsymbol{I}$ to $\bar{\boldsymbol{\Omega}}$ with κ picked to make minimum eigenvalue of $\boldsymbol{\Omega} = \bar{\boldsymbol{\Omega}} + \kappa \boldsymbol{I}$ equal to 0.5. Let $\boldsymbol{\Omega} = \tilde{\boldsymbol{F}}\tilde{\boldsymbol{F}}$ (matrix square-root), then $\boldsymbol{F} = \tilde{\boldsymbol{F}}^{-1}$ in (3).

We applied our proposed approach with $n = 64$ or 256, $m = 4$ (maximum delay 3), and compared with the approach of [17] ($M = 2$, $K = 15$ for $n = 64$ and $K = 63$ for $n = 256$) and the approach of [5] (which is also the approach of [6,7], all of whom assume i.i.d. observations and have two lasso penalties one each on $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$, counterpart to our $\tilde{\boldsymbol{\Gamma}}$ with no delays). By changing the penalty parameters and determining the resulting edges, we calculated the true positive rate (TPR) and false positive rate 1-TNR (where TNR is the true negative rate) over 100 runs, separately for $\boldsymbol{\Omega}$ and $\tilde{\boldsymbol{\Gamma}}/\{\boldsymbol{\Phi}_k\}/\boldsymbol{\Gamma}$ ($\{\boldsymbol{\Phi}_k\}$ are the inverse PSD's in [17]). The receiver operating characteristic (ROC) is shown in Figs. 1 and 2 based on 100 runs. Figs. 1- 2 show that the i.i.d. modeling of [5–7] is unable to capture the "dependent" edges (cf. (3)) via $\boldsymbol{\Gamma}$ whereas it has no issues with $\boldsymbol{\Omega}$. Our embedding approach as well as the frequency-domain approach of [17], both work well for both components of the graph Kronecker product, with our embedding approach being better (higher TPR for a given 1-TNR).

In Table 1, we compare the three approaches in terms of the F_1 score, execution time (based on tic-toc functions in MATLAB), TPR and 1-TNR, for fixed penalty parameter λ selected from a grid of values (the same as for computing the ROC curves) to maximize the F_1 score averaged over 100 runs. It is seen that the delay embedding approach is faster than the frequency-domain approach.

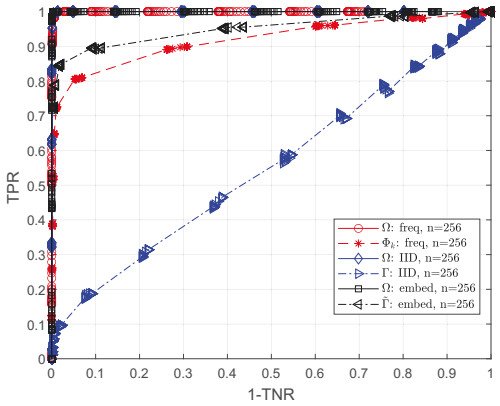

Fig. 2: ROC curves for $n = 256$. Other description as for Fig. 1.

7. CONCLUSIONS

Inference of the conditional independence graph of a sparse, high-dimensional, stationary matrix-variate Gaussian time series was considered under the assumption that the correlation function of the matrix series is Kronecker-decomposable. A time-delay embedding approach was proposed where with each matrix node, we associate a random vector consisting of a scalar series component and its time-delayed copies. A group-lasso penalized negative pseudo log-likelihood (NPLL) objective function was formulated and optimized via flip-flop minimization. We illustrated our approach using a numerical example where our approach significantly outperformed an existing i.i.d. modeling-based approach [5–7] as well as an existing frequency-domain approach [17] for dependent data, in correctly detecting the graph edges with ROC as the performance metric.

Future work includes performance analysis and application to real data.

8. REFERENCES

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

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

[3] 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.

[4] 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.

[5] C. Leng and C.Y. Tang, "Sparse matrix graphical models," *J. American Statistical Association*, vol. 107, pp. 1187-1200, Sep. 2012.

[6] J. Yin and H. Li, "Model selection and estimation in the matrix normal graphical model," *J. Multivariate Analysis*, vol. 107, pp. 119-140, May 2012.

[7] T. Tsiligkaridis, A.O. Hero, III, and S. Zhou, "On convergence of Kronecker graphical lasso algorithms," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1743-1755, April 2013.

[8] S. Zhou, "Gemini: Graph estimation with matrix variate normal instances," *Annals Statistics*, vol. 42, no. 2, pp. 532-562, 2014.

[9] S. He, J. Yin, H. Li and X. Wang, "Graphical model selection and estimation for high dimensional tensor data," *J. Multivariate Analysis*, vol. 128, pp. 165-185, 2014.

[10] F. Huang and S. Chen, "Joint learning of multiple sparse matrix Gaussian graphical models," *IEEE Trans. Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 2606-2620, Nov. 2015.

[11] Y. Zhu and L. Li, "Multiple matrix Gaussian graphs estimation," *J. Royal Statistical Society, Series B (Methodological)*, vol. 80, pp. 927-950, 2018.

[12] X. Chen and W. Liu, "Graph estimation for matrix-variate Gaussian data," *Statistica Sinica*, vol. 29, pp. 479-504, 2019.

[13] K. Greenewald, S. Zhou and A. Hero III, "Tensor graphical lasso (teralasso)," *J. Royal Statistical Society, Series B (Methodological)*, vol. 81, no. 5, pp. 901-931, 2019.

[14] X. Lyu, W.W. Sun, Z. Wang, H. Liu, J. Yang and G. Cheng, "Tensor graphical model: Non-convex optimization and statistical inference," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2024-2037, 1 Aug. 2020.

[15] A.K. Gupta and D.K. Nagar, *Matrix Variate Distributions*. Boca Raton, FL: Chapman and Hall/CRC Press, 1999.

[16] K. Werner, M. Jansson and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 478-491, Feb. 2008.

[17] J.K. Tugnait, "Sparse high-dimensional matrix-valued graphical model learning from dependent data," in *Proc. 2023 IEEE Statistical Signal Processing Workshop (SSP-2023)*, Hanoi, Vietnam, July 2-5, 2023, pp. 344-348.

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

[19] 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, pp. 4508-4512.

[20] F. Takens, "Detecting strange attractors in turbulence," *Dynamical Systems and Turbulence*, vol. 898, no. 1, pp. 365-381, 1981.

[21] J. Frank, S. Mannor and D. Precup, "Activity and gait recognition with time-delay embeddings," in *AAAI'10: Proc, Twenty-Fourth AAAI Conf. Artificial Intelligence*, AAAI Press, Atlanta, GA, July 2010, pp. 1581-1586.

[22] S.M. Hirsch, S.M. Ichinaga, S.L. Brunton, J.N. Kutz and B.W. Brunton, "Structured time-delay models for dynamical systems with connections to Frenet-Serret frame," *Proc. Royal Soc. A: Mathematical, Physical and Engineering Sciences*, vol. 477, no. 2254, pp. 20210097, 2021.

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

[24] J. Gorski, F. Pfeuffer and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Mathematical Methods of Operations Research*, vol. 66, pp. 373-408, 2007.

[25] 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.

[26] 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.)