Estimation of a causal directed acyclic graph process using non-gaussianity

Aref Einizade, Jhony H. Giraldo, Fragkiskos D. Malliaros, Sepideh Hajipour Sardouie

Estimation of a Causal Directed Acyclic Graph Process using Non-Gaussianity

Aref Einizade, Jhony H. Giraldo, Fragkiskos D. Malliaros, and Sepideh Hajipour Sardouie,

***Abstract*—Numerous approaches have been proposed to discover causal dependencies in machine learning and data mining; among them, the state-of-the-art VAR-LiNGAM (short for Vector Auto-Regressive Linear Non-Gaussian Acyclic Model) is a desirable approach to reveal *both* the instantaneous and time-lagged relationships. However, all the obtained VAR matrices need to be analyzed to infer the final causal graph, leading to a rise in the number of parameters. To address this issue, we propose the CGP-LiNGAM (short for Causal Graph Process-LiNGAM) method, which has significantly fewer model parameters and deals with only *one* causal graph, i.e., directed acyclic graph (DAG), for interpreting the causal relations by exploiting Graph Signal Processing (GSP). Besides, the Graph Shift Invariance and Uniqueness of the proposed approach are also studied and shown. Experimental results demonstrate the superiority and robustness over the state-of-the-art methods, especially in a high amount of noise. Besides, the real-world applicability and interpretability of the proposed approach are shown in learning effective brain connectivity during sleep, and the compatibility of obtained causal directed brain graphs with the previous sleep-related neuroscientific studies is investigated.**

***Index Terms*—Causal Discovery, Graph Signal Processing (GSP), Causal Graph Process (CGP), Linear Non-Gaussian Acyclic Model (LiNGAM), Directed Acyclic Graph (DAG).**

I. Introduction

Causal Discovery (CD) [1], [2] of time-series data, also known as Dynamic CD (DCD), has gained significant attention due to its capability to reveal causal relationships between observational data. A wide variety of DCD methods have been proposed based on different points of view. For a comprehensive survey, please refer to [3] and the references therein. Two popular and well-known types of causal dependency order are instantaneous and time-lagged ones which are mostly addressed by Structural Equation Model (SEM) [2], and Vector Auto-Regressive (VAR) ones [2], respectively. Among the SEM-based approaches, the popular Linear Non-Gaussian Acyclic Model (LiNGAM) addresses the identifiability issue by assuming the exogenous disturbances are non-Gaussian [4]. Besides, the state-of-the-art VAR-LiNGAM method [5] has been proposed to reveal both the instantaneous and time-lagged dependencies. As well as the desirable advantages of the VAR and VAR-LiNGAM, having a high number of free model parameters and the need for interpreting all VAR matrix coefficients to infer the final causal graph are still major drawbacks [6].

Aref Einizade (Corresponding author, email: aref.einizade@yahoo.com), and Sepideh Hajipour Sardouie are with the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran. Jhony H. Giraldo is with the LTCI, Telecom Paris - Institut Polytechnique de Paris, France. Fragkiskos D. Malliaros is with the Universite Paris-Saclay, CentraleSupelec, Inria, Centre for Visual Computing (CVN), France.

On the other hand, in most recording systems of time series, the interacting sensors, such as brain regions [7], geographic temperature locations [6], etc., can be considered the nodes of a meaningful underlying (and possibly directed) causal graph. For example, the learned graphs from brain signals can reveal the directed interactions between latent brain sources [5], [8], [9]. These graphs are rarely known in real machine learning applications and need to be carefully learned/estimated from observational data [3], [7]. Building upon this causal graph, the different recorded measurements on different nodes (in *one* time index) can be interpreted as a graph signal, which provides utilizing Graph Signal Processing (GSP) tools [10], [11], [6], [12] in different areas such as image [13], [14], [15] and video [16], [17] processing.

Recently, a GSP-based method named Causal Graph Process (CGP) [6] has been proposed to address the mentioned drawbacks of the VAR by modeling the VAR coefficient matrices with graph polynomial filters; however, this model can not model the instantaneous dependencies. To address the mentioned issues, in the proposed (graph) shift invariant [6] CGP-LiNGAM analysis, the discovery of causal relationships relies on only *one* and *shared* underlying Directed Acyclic Graph (DAG) [6] to reveal both the instantaneous and time-lagged dependencies, unlike almost all of the other VAR-based approaches, such as VAR-LiNGAM [5] and DYNOTEARS [18], [19], [20], which need analyzing all the obtained VAR matrix coefficients to infer the underlying causal graph.

From the viewpoint of real-world applicability, the GSP tools have gained great attention in a wide range of applications, especially brain imaging and analysis [21]. The flexibility of GSP concepts and tools to describe direct or indirect connections makes it a powerful choice for processing brain signals or inferring brain connectivity, which plays an important role in clinical diagnosis, especially during sleep [22]. However, to the best of our knowledge, the inference of sleep-related effective brain connectivity from a GSP point of view has not been addressed, and the current work aims to investigate it.

The main contributions of the present paper are summarized as follows:

- The proposed CGP-LiNGAM reveal both the instantaneous and time-lagged dependencies and is related to only *one* and *shared* underlying Directed Acyclic Graph (DAG) [6] (significantly fewer parameters, based on the analysis presented in Section (III-D)) unlike the state-of-the-art VAR-LiNGAM [5] and DYNOTEARS [20].
- We present closed-form iteration-based solutions in Section III, which facilitate applicability and reproducibility.

- The proposed model has the desired property of Graph Shift Invariance (GSI) [23] as will be shown in Theorem 2. Besides, the non-Gaussianity of the exogenous disturbances guarantees the identifiability of the proposed model [5], as will be proved in Theorem 3.
- Based on the experimental analysis in Sections IV-A, IV-B and IV-C, the CGP-LiNGAM performs more accurately and robustly in causal graph recovery and prediction tasks compared to the VAR-LiNGAM [5], and DYNOTEARS [20].
- In Section IV-D, the efficiency of the proposed approach over both the generic LiNGAM [4] and CGP [6] in handling both the instantaneous and lagged causal effects is illustrated.
- Based on the additional presented analysis in Section IV-E, the CGP-LiNGAM method is also robust in the case of not known CGP (VAR) order [24].
- In addition to the mentioned numerical analysis, the learning effective brain connectivity during sleep [22], [8] will be investigated in Section IV-F to show the real-world applicability and neuroscientific compatibility of the inferred brain graphs.

The notations $(.)^\top$, $\otimes$, $(.)^\dagger$, $\|.\|_p$, and $\hat{x}$ stand for the transpose operator, Kronecker product, Moore-Penrose pseudo-inverse of a matrix, the p-norm of a vector, and the estimation of the true entity x, respectively.

II. Background and Related Work

The observation matrix $\mathbf{X} = (\mathbf{x}(0), \mathbf{x}(1), ..., \mathbf{x}(K-1)) \in \mathbb{R}^{N\times K}$ is provided to us by recording K-length signals on N sensors/nodes. Firstly, the GSP background is briefly outlined.

A. GSP background

A directed graph $\mathcal{G} = (\mathbf{V}, \mathbf{A})$ is characterized by its vertex set $\mathbf{V} = \{v_1, ..., v_N\}$ and its adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$. If $\mathbf{A}_{ij} \neq 0$, the ith node is a parent of the jth node, and the jth node is a descendant of the ith node [2]. We call a node *pure parent* if it is not the descendant of any other nodes. A graph signal $\mathbf{x} = (x_1, ..., x_N)^\top \in \mathbb{R}^{N\times 1}$ is a mapping $\mathbf{x} : \mathbf{V} \to \mathbb{R}$, assigning the ith vertex the value of x_i [10]. A L-order graph polynomial filter is also defined as [6]:

$$P(\mathbf{A}, \mathbf{c}) = c_0\mathbf{I} + c_1\mathbf{A} + ... + c_L\mathbf{A}^L, \tag{1}$$

where $\mathbf{I}$ denotes the identity matrix of size N, and $\mathbf{c} = (c_0, ..., c_L)^\top$ contains the scalar filter coefficients.

In the following, the related work to the current study is briefly described.

B. LiNGAM

In an (instantaneous w.r.t. the current time sample k, where $k = 0, ..., K-1$) SEM (2), assuming the exogenous disturbance $\mathbf{e}$ is non-Gaussian, the underlying DAG $\mathbf{A}$ can be *uniquely* recovered by the LiNGAM analysis [4].

$$\mathbf{x}(k) = \mathbf{A}\mathbf{x}(k) + \mathbf{e}(k). \tag{2}$$

C. VAR

The VAR modeling of time-series $\mathbf{X}$ for $k = 0, ..., K-1$ is defined as:

$$\mathbf{x}(k) = \sum_{i=1}^{M} \mathbf{R}_i\mathbf{x}(k-i) + \mathbf{e}(k), \tag{3}$$

where M denotes the VAR order, and $\{\mathbf{R}_i\}_{i=1}^M$ are VAR coefficient matrices [5]. From the Granger [25] analysis point of view, to recover the underlying causal graph $\mathbf{A}'$ modeling the causal relationships, one considers $\mathbf{A}'_{ij} = 0$, if $\mathbf{R}_k^{(ij)} = 0$ for all k [26], [6], implying the conditional independence between the time-series at each node described by a Markov Random Field (MRF) with adjacency structure $\mathbf{A}'$ [6], where $\mathbf{A}'_{ij}$ and $\mathbf{R}_k^{(ij)}$ denote the (i,j)th element of $\mathbf{A}'$ and $\mathbf{R}_k$, respectively.

D. VAR-LiNGAM

In this method, both the instantaneous and time-lagged dependencies are modeled using the combination of (2) and (3) as [5]:

$$\mathbf{x}(k) = \sum_{i=0}^{M} \mathbf{R}_i\mathbf{x}(k-i) + \mathbf{e}(k). \tag{4}$$

In this method, $\mathbf{R}_0$ is considered the adjacency matrix of an underlying DAG, and the exogenous disturbance $\mathbf{e}$ is assumed to be non-Gaussian to guarantee the identifiability [5]. Then, similar to the VAR model (3), the sparsity pattern of all $\{\mathbf{R}_i\}_{i=0}^M$ must be analyzed to recover the true underlying causal graph $\mathbf{A}'$ [5].

E. Causal Graph Process (CGP)

An M-order CGP model has the following form [6]:

$$\mathbf{x}(k) = \sum_{i=1}^{M} P_i(\mathbf{A}, \mathbf{c})\mathbf{x}(k-i) + \mathbf{e}(k), \tag{5}$$

where $P_i(\mathbf{A}, \mathbf{c}) = \sum_{j=0}^{i} c_{ij}\mathbf{A}^j$ implements a graph filtering operator and $\mathbf{c}$ collects the scalar polynomial coefficients as ($1 \leq i \leq M$ and $0 \leq j \leq i$):

$$\mathbf{c} = (c_{10}, c_{11}, ..., c_{ij}, ..., c_{MM})^\top. \tag{6}$$

III. The Proposed CGP-LiNGAM Approach

In our CGP-LiNGAM framework, the instantaneous (LiNGAM) and time-lagged (CGP) parts share *one* underlying DAG adjacency matrix $\mathbf{A}$. To recover both these causal dependencies, our proposed framework is shown as follows:

$$\mathbf{x}(k) = \mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{M} \overbrace{P_i(\mathbf{A}, \mathbf{c})}^{\mathbf{R}_i}\mathbf{x}(k-i) + \mathbf{e}(k). \tag{7}$$

The proposed model (7) can be rewritten as:

$$\mathbf{x}(k) = \sum_{i=1}^{M} \tilde{\mathbf{R}}_i\mathbf{x}(k-i) + \tilde{\mathbf{e}}(k), \tag{8}$$

where the following theorem helps to recover $\tilde{\mathbf{R}}_i$ in (8), and

$$\tilde{\mathbf{R}}_i = (\mathbf{I}-\mathbf{A})^{-1}\mathbf{R}_i;\ \tilde{\mathbf{e}}(k) = (\mathbf{I}-\mathbf{A})^{-1}\mathbf{e}(k). \tag{9}$$

The following theorem shows that how the pair matrices $\tilde{\mathbf{R}}_i$ and $\tilde{\mathbf{R}}_j$ commute, for $i,j = 1,2,...,M$.

Theorem 1. *(Commutativity). In the model (8), given that $\boldsymbol{A}$ is the adjacency matrix of a DAG, $\tilde{\boldsymbol{R}}_i$ and $\tilde{\boldsymbol{R}}_j$ commute, i.e., $\tilde{\boldsymbol{R}}_i\tilde{\boldsymbol{R}}_j = \tilde{\boldsymbol{R}}_j\tilde{\boldsymbol{R}}_i$, for $i,j = 1,2,...,M$.*

Proof: If $\mathbf{A}$ is the adjacency matrix of a DAG, $\mathbf{I}-\mathbf{A}$ is invertible [4], [5] and also $\forall i:\ |\lambda_i(\mathbf{A})| < 1$. In fact, due to the strict lower triangularity of $\mathbf{A}$, $\forall i:\ \lambda_i(\mathbf{A}) = 0$. Therefore, the infinite sum of the geometric series $\sum_{r=0}^{\infty}\mathbf{A}^r$ is summable, and equal to $(\mathbf{I}-\mathbf{A})^{-1}$ as [27]:

$$(\mathbf{I}-\mathbf{A})^{-1} = \sum_{r=0}^{\infty}\mathbf{A}^r. \tag{10}$$

Then, the VAR coefficient matrices $\{\tilde{\mathbf{R}}_i\}_{i=1}^M$ in (8) commute because they can be stated as (infinite) graph polynomial filters:

$$\begin{aligned}\tilde{\mathbf{R}}_i = (\mathbf{I}-\mathbf{A})^{-1}\mathbf{R}_i &= \left[\sum_{r=0}^{\infty}\mathbf{A}^r\right]\left[\sum_{j=0}^{i}c_{ij}\mathbf{A}^j\right]\\ &= \sum_{r=0}^{\infty}\sum_{j=0}^{i}c_{ij}\mathbf{A}^{r+j}. \qquad \blacksquare\end{aligned} \tag{11}$$

The next theorem outlines the property of the Graph Shift Invariance (GSI) for the proposed CGP-LiNGAM model (7).

Theorem 2. *(Graph Shift Invariance). In the proposed CGP-LiNGAM model (7), if the new input $\boldsymbol{x}'$ is the graph filtered version of the initial input $\boldsymbol{x}$ induced by the graph filter $h_{\boldsymbol{c}'}(\boldsymbol{A}) = P(\boldsymbol{A},\boldsymbol{c}')$, i.e., $\boldsymbol{x}'(k) = h_{\boldsymbol{c}'}(\boldsymbol{A})\boldsymbol{x}(k);\ k = 1,...,K$, the new input $\boldsymbol{x}'$ still can be described by the CGP-LiNGAM model (7) with the same parameters $\{\boldsymbol{A},\boldsymbol{c}\}$.*

proof: Firstly, to use brief notations in the proof, we denote $h_{\mathbf{c}}(\mathbf{A}) = P(\mathbf{A},\mathbf{c})$. Then, due to the commutativity of two arbitrary graph filters $h_{\mathbf{c}}(\mathbf{A})$ and $h_{\mathbf{c}'}(\mathbf{A})$ [6] and also since the multiplication of $h_{\mathbf{c}}(\mathbf{A})$ and $h_{\mathbf{c}'}(\mathbf{A})$ is a new graph filter $h_{\mathbf{c}''}(\mathbf{A})$ [23], based on the definition of $\mathbf{x}'$, one can write:

$$\begin{aligned}\mathbf{x}'(k) &= h_{\mathbf{c}'}(\mathbf{A})\mathbf{x}(k)\\ &= h_{\mathbf{c}'}(\mathbf{A})\left[\mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{M}h_{\mathbf{c}_i}(\mathbf{A})\mathbf{x}(k-i) + \mathbf{e}(k)\right]\\ &= [h_{\mathbf{c}'}(\mathbf{A})\mathbf{A}]\mathbf{x}(k) + \sum_{i=1}^{M}[h_{\mathbf{c}'}(\mathbf{A})h_{\mathbf{c}_i}(\mathbf{A})]\mathbf{x}(k-i) + \overbrace{h_{\mathbf{c}'}(\mathbf{A})\mathbf{e}(k)}^{\mathbf{e}'(k)}\\ &= \mathbf{A}[h_{\mathbf{c}'}(\mathbf{A})\mathbf{x}(k)] + \sum_{i=1}^{M}h_{\mathbf{c}_i}(\mathbf{A})[h_{\mathbf{c}'}(\mathbf{A})\mathbf{x}(k-i)] + \mathbf{e}'(k)\\ &= \mathbf{A}\mathbf{x}'(k) + \sum_{i=1}^{M}h_{\mathbf{c}_i}(\mathbf{A})\mathbf{x}'(k-i) + \mathbf{e}'(k).\end{aligned} \tag{12}$$

Therefore, the new input $\mathbf{x}'$ still can be described by the CGP-LiNGAM model (7) with the same underlying DAG $\mathbf{A}$ and graph filter matrix coefficients $\{h_{\mathbf{c}_i}(\mathbf{A})\}_{i=1}^M$. $\blacksquare$

The proposed CGP-LiNGAM algorithm is summarized in Algorithm 1, but in the following, the proposed three-step approach for recovering $\{\tilde{\mathbf{R}}_i\}_{i=1}^M$, $\mathbf{A}$, $\{\mathbf{R}_i\}_{i=1}^M$, and $\mathbf{c}$ is presented in detail. However, from the proposed CGP-LiNGAM model (7), it can be seen that the main goal is recovering the model parameters $\{\mathbf{A},\mathbf{c}\}$, and the intermediate variables $\{\tilde{\mathbf{R}}_i\}_{i=1}^M$ and $\{\mathbf{R}_i\}_{i=1}^M$ are described only for the sake of clarity.

A. Recovering $\{\tilde{\boldsymbol{R}}_i\}_{i=1}^M$

Based on Theorem 1, the commutativity terms are included in the following multi-convex optimization [6] to obtain the (possibly sparse) graph polynomial filters $\left\{\tilde{\mathbf{R}}_i\right\}_{i=1}^M$ as:

$$\begin{aligned}&\hat{\tilde{\mathbf{R}}}_i = \underset{\tilde{\mathbf{R}}_i}{\operatorname{argmin}}\ \frac{1}{2}\sum_{k=M}^{K-1}\left\|\mathbf{x}(k) - \sum_{i=1}^{M}\tilde{\mathbf{R}}_i\mathbf{x}(k-i)\right\|_2^2 + \lambda_1\|vec(\tilde{\mathbf{R}}_1)\|_1\\ &\text{Subject to:}\ \ \tilde{\mathbf{R}}_i\tilde{\mathbf{R}}_j = \tilde{\mathbf{R}}_j\tilde{\mathbf{R}}_i;\quad i,j = 1,...,M.\end{aligned} \tag{13}$$

An alternating optimization to (13) is expressed as:

$$\begin{aligned}\hat{\tilde{\mathbf{R}}}_i = \underset{\tilde{\mathbf{R}}_i}{\operatorname{argmin}}\ &\frac{1}{2}\sum_{k=M}^{K-1}\left\|\mathbf{x}(k) - \sum_{i=1}^{M}\tilde{\mathbf{R}}_i\mathbf{x}(k-i)\right\|_2^2\\ &+ \lambda_1\|vec(\tilde{\mathbf{R}}_1)\|_1 + \lambda_3\sum_{i\neq j}\|\tilde{\mathbf{R}}_i\tilde{\mathbf{R}}_j - \tilde{\mathbf{R}}_j\tilde{\mathbf{R}}_i\|_F^2,\end{aligned} \tag{14}$$

where, with the definitions

$$\mathbf{X}_m = (\mathbf{x}(m),\ \mathbf{x}(m+1),\ ...,\ \mathbf{x}(m+K-M-1)), \tag{15}$$

and

$$\tilde{\mathbf{r}}_i = vec(\tilde{\mathbf{R}}_i), \tag{16}$$

(14) can be rewritten as (with the details [28] in the Appendix (Section VI)):

$$\hat{\tilde{\mathbf{r}}}_i = \underset{\tilde{\mathbf{r}}_i}{\operatorname{argmin}}\ \|\boldsymbol{\Psi}_i\tilde{\mathbf{r}}_i - \tilde{\mathbf{y}}_i\|_2^2 + \lambda_1\|\tilde{\mathbf{r}}_1\|_1, \tag{17}$$

where $\mathbf{0}$ is a all-zero array with appropriate size, and

$$\boldsymbol{\Psi}_i = \left(\frac{\sqrt{2}}{2}\mathbf{B}_i^\top, \sqrt{\lambda_3}\boldsymbol{\Phi}_1^\top, .., \sqrt{\lambda_3}\boldsymbol{\Phi}_{i-1}^\top, \sqrt{\lambda_3}\boldsymbol{\Phi}_{i+1}^\top, .., \sqrt{\lambda_3}\boldsymbol{\Phi}_M^\top\right)^\top, \tag{18}$$

$$\tilde{\mathbf{y}}_i = \left(\frac{\sqrt{2}}{2}\mathbf{y}_i^\top, \mathbf{0}^\top, \ldots, \mathbf{0}^\top\right)^\top, \tag{19}$$

$$\begin{aligned}&\mathbf{B}_i = \mathbf{X}_{M-i}^\top \otimes \mathbf{I},\quad \boldsymbol{\Phi}_i = \mathbf{R}_i^\top \otimes \mathbf{I} - \mathbf{I}^\top \otimes \mathbf{R}_i,\\ &\mathbf{y}_i = vec(\mathbf{X}_M) - vec\left(\sum_{j=1\neq i}^{M}\tilde{\mathbf{R}}_j\mathbf{X}_{M-j}\right).\end{aligned} \tag{20}$$

Algorithm 1 : CGP-LiNGAM

Input: $\mathbf{X} \in \mathbb{R}^{N\times T}$, M, $\{\lambda_i\}_{i=1}^{3}$
Output: Causal DAG $\mathbf{A} \in \mathbb{R}^{N\times N}$, Polynomial coefficients $\mathbf{c}$

1: Initialize: the iteration index $t = 1$, $\hat{\mathbf{R}}^{(0)} = \mathbf{0}$
2: **while** Convergence **do**
3: **for** $i = 1 : M$ **do**
4: Estimate $\tilde{\mathbf{R}}_i^{(t)}$ by fixing $\{\tilde{\mathbf{R}}_j^{(t-1)}\}_{j=1\neq i}^{M}$, and using CVX [31] in (13) or (14), or using LASSO[29]/ closed-form solutions in (21)
5: **end for**
6: $t \leftarrow t + 1$
7: **end while**: return $\{\tilde{\mathbf{R}}_i\}_{i=1}^{M}$ and $\tilde{\mathbf{e}}$ in (8)
8: Estimate $\mathbf{R}_0 = \mathbf{A}$ using LiNGAM analysis [4] on $\tilde{\mathbf{e}}$
9: Obtain the causal effect matrices $\{\mathbf{R}_i = (\mathbf{I} - \mathbf{A})\tilde{\mathbf{R}}_i\}_{i=1}^{M}$
10: Solve for $\mathbf{c}$ in (23) by l-1 minimization, e.g., LASSO [29]

Afterwards, the *unique* (identifiable) [6] closed-form solutions to (14) are expressed as:

$$\hat{\tilde{\mathbf{r}}}_{i\neq 1} = \mathbf{\Psi}_i^{\dagger}\tilde{\mathbf{y}}_i, \quad i = 2, ..., M, \tag{21}$$

and $\hat{\tilde{\mathbf{r}}}_1$ can be obtained using standard l-1 minimization approaches in (17), such as LASSO [29]. Note that the proposed steps of (17) and (21) for $i = 1, ..., M$ are repeated until convergence (please refer to Lines 2-7 in Algorithm 1).

B. Recovering $\boldsymbol{A}$

By recovering $\{\tilde{\mathbf{R}}_i\}_{i=1}^{M}$, the residual $\tilde{\mathbf{e}}$ is also obtained from (8), and the relation $\tilde{\mathbf{e}}(k) = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{e}(k)$ can be rewritten:

$$\tilde{\mathbf{e}}(k) = \mathbf{A}\tilde{\mathbf{e}}(k) + \mathbf{e}(k), \tag{22}$$

which is an SEM, and, due to the non-Gaussianity of the exogenous disturbance $\mathbf{e}$, the underlying DAG $\mathbf{A}$ can be *uniquely* (identifiable) inferred using the LiNGAM analysis [4], [5] on $\tilde{\mathbf{e}}$ exploiting FastICA [30] approach as will be shown in Theorem 3.

C. Recovering $\{\boldsymbol{R}_i\}_{i=1}^{M}$ and $\boldsymbol{c}$

By recovering $\{\tilde{\mathbf{R}}_i\}_{i=1}^{M}$ and $\mathbf{A}$, the graph polynomial filters $\{\mathbf{R}_i = (\mathbf{I} - \mathbf{A})\tilde{\mathbf{R}}_i\}_{i=1}^{M}$ are obtained, and, inspired from the CGP model [6] (i.e., $\mathbf{R}_i = P_i(\mathbf{A}, \mathbf{c}) = \sum_{j=0}^{i} c_{ij}\mathbf{A}^j$), the polynomial coefficients $\mathbf{c}$ are estimated by minimizing the following convex optimization using standard l_1-regularized least squares methods, such as LASSO [29]:

$$\hat{\mathbf{c}}_i = \underset{\mathbf{c}_i}{\operatorname{argmin}} \frac{1}{2}\left\|vec(\hat{\mathbf{R}}_i) - \mathbf{Q}_i\mathbf{c}_i\right\|_2^2 + \lambda_2\|\mathbf{c}_i\|_1, \tag{23}$$

where $\mathbf{Q}_i = \left(vec(\mathbf{I}), vec(\hat{\mathbf{A}}), ..., vec(\hat{\mathbf{A}}^i)\right)$, and $\mathbf{c}_i = (c_{i0}, c_{i1}, ..., c_{ii})^{\top}$ [6] for $i = 1, ..., M$. Concisely, our proposed CGP-LinGAM algorithm is summarized in Algorithm 1. The following theorem describes the needed mild assumptions for the CGP-LiNGAM (7) to be uniquely recovered.

Theorem 3. *(Uniqueness). In the CGP-LiNGAM (7) with known AR order M, given that $\boldsymbol{A}$ is the adjacency matrix of an underlying DAG, the exogenous noise $\boldsymbol{e}$ is non-Gaussian, and under the mild conditions $c_{10} = 0$ and $c_{11} = 1$, the model parameters $\{\boldsymbol{A}, \{\boldsymbol{c}_i\}_{i=1}^{M}\}$ can be uniquely recovered.*

proof: It has been shown [6] that, under very mild conditions, i.e., $c_{10} = 0$ and $c_{11} = 1$, in the CGP model (5), $\mathbf{A}$ and $\mathbf{c}$ can be uniquely described without ambiguity. Therefore, with multi-convexity of (13), the only remaining step for checking uniqueness and identifiability is the LiNGAM step (22). In this step, precisely, a classic Independent Component Analysis (ICA) method, i.e., FastICA in LiNGAM [4], is used, which, under the Non-Gaussianity of the exogenous noise $\mathbf{e}$, the obtain DAG $\mathbf{A}$ is unique and identifiable [4]. ■

D. Number of learnable model parameters

In the CGP-LiNGAM (7), the underlying DAG $\mathbf{A}$ has $\frac{N(N-1)}{2}$ free parameters (because of the technically being lower triangular), and $\mathbf{c} \in \mathbb{R}^{\frac{M(M+3)}{2}}$, compared to the VAR-LiNGAM [5] (4), which needs $\frac{N(N-1)}{2} + MN^2$ free parameters to describe the model. On the other hand, in real applications, usually $N \gg M$ [5], [6], so the CGP-LiNGAM (with combined $\frac{N(N-1)}{2} + \frac{M(M+3)}{2}$ free parameters) is considerably parsimonious compared to the VAR-LiNGAM.

E. Complexity Analysis of CGP-LiNGAM (Algorithm 1)

For a specific i and an convergence iteration in Line 4, optimization (14) is naively dominated by the matrix-matrix product of $\{\tilde{\mathbf{R}}_i\tilde{\mathbf{R}}_j\}_{j=1\neq i}^{M}$ with $\mathcal{O}((M-1)N^3)$, and matrix-vector product of $\{\mathbf{R}_j\mathbf{x}(k-j)\}_{k=M,j=1}^{K-1,M}$ with $\mathcal{O}((K-M)N^2)$ complexity [6]. As a result, the overall complexity of one convergence iteration is $\mathcal{O}(M^2N^3 + KMN^2)$. In Line 8, the computational complexity of the LiNGAM analysis typically $\mathcal{O}(KN^3 + N^4)$ [32]. The matrix-matrix products of $\{(\mathbf{I} - \mathbf{A})\tilde{\mathbf{R}}_i\}_{i=1}^{M}$ in Line 9 take total complexity of $\mathcal{O}(MN^3)$. In Line 10, the minimization (23) for each i is dominated by multiplications $\mathbf{Q}_i^{\top}\mathbf{Q}_i \in \mathbb{R}^{(i+1)\times(i+1)}$ and $\mathbf{Q}_i^{\top}vec(\tilde{\mathbf{R}}_i) \in \mathbb{R}^{(i+1)\times 1}$, which take $\mathcal{O}((i+1)^2)$ operations [6], and, combined complexity of $\mathcal{O}(M(\sum_{i=1}^{M}(i+1)^2)) \approx \mathcal{O}(M^4)$. All in all, assuming $M \ll K$, the total (worst-case and naive) computational complexity of Algorithm 1 is approximately $\mathcal{O}(KMN^2 + KN^3 + N^4 + M^4)$. Note that the (possible) sparsity of the underlying DAG can severely reduce the complexity [6].

IV. Experimental Results and Discussion

In this section, we provide experimental analysis in two categories: 1) ablation study: analyzing the performance of the CGP-LiNGAM in different situations, 2) comparison study: comparing the performance of the CGP-LiNGAM with that of the state-of-the-art, i.e., VAR-LiNGAM [5]. The underlying DAG $\mathbf{A}$ and also the non-Gaussian exogenous disturbance $\mathbf{e}$ are generated as described in [4]. Note that, to have sparse DAGs, we generate $\mathbf{A}$ (with $N = 5$) with just *one* pure parent. The generated data were divided into three parts of the train, validation (to find the optimal hyperparameters using

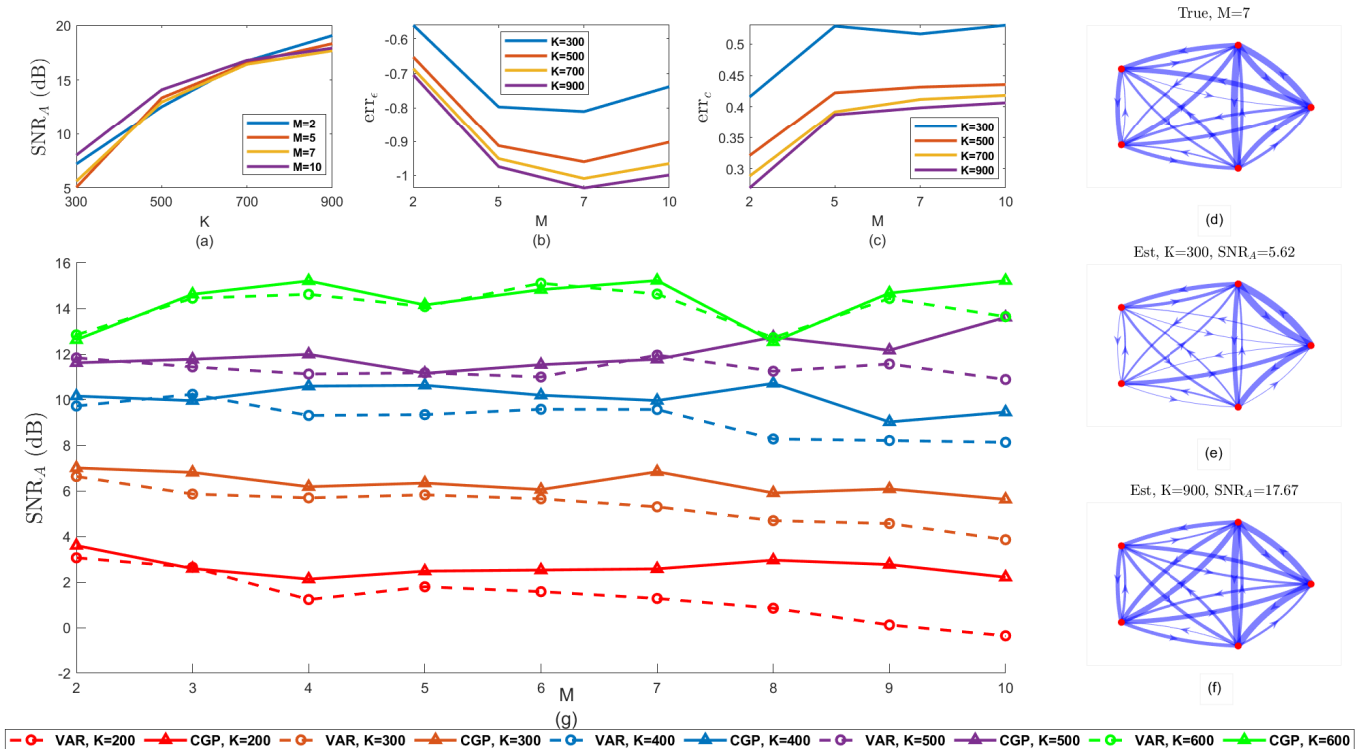

Fig. 1: (a-c) The average of the performance evaluation metrics over thirty Monte-Carlo realizations, by varying the number of samples K and the CGP-order M in the spans of $\{300, 500, 700, 900\}$ and $\{2, 5, 7, 10\}$, respectively. (d-f): An example average ($M = 7$) to the graph structures (Figure 1, (d:f)) over Monte-Carlo realizations. (g): Comparison of the graph recovery performance of the proposed CGP-LiNGAM method with the celebrated VAR-LiNGAM [5]

grid search) and test (to evaluate the prediction based on the optimized model) [6].

A. Ablation Study

We vary the number of samples K and the CGP-order M in the spans of $\{300, 500, 700, 900\}$ and $\{2, 5, 7, 10\}$, respectively, to study the effect of these changes on the recovery performance. The polynomial coefficients $\mathbf{c}$ are generated as $2^{i+j}c_{ij} \sim 0.5[\mathcal{U}(-1, -0.45) + \mathcal{U}(0.45, 1)]$ to model coefficient decay with distance to the current time sample [6]. Besides, to evaluate the quality of the recoveries, we consider the following scale-free metrics $SNR_{\mathbf{A}} = 20\log(\|\mathbf{A}\|_F/\|\hat{\mathbf{A}} - \mathbf{A}\|_F)$, $err_{\mathbf{c}} = \|\hat{\mathbf{c}} - \mathbf{c}\|_2/\|\mathbf{c}\|_2$, and

$$err_\epsilon = \mathbb{E}\left[\frac{1}{N}\|\mathbf{x}(k) - f(\hat{\mathbf{A}}, \hat{\mathbf{c}}, \mathbf{X}'_{k-1})\|_2^2\right] - \mathbb{E}\left[\frac{1}{N}\|\mathbf{x}(k) - f(\mathbf{A}, \mathbf{c}, \mathbf{X}'_{k-1})\|_2^2\right], \tag{24}$$

where $f(\mathbf{A}, \mathbf{c}, \mathbf{X}'_{k-1}) = \mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{M} P_i(\mathbf{A}, \mathbf{c})\mathbf{x}(k-i)$, and err_ϵ models the prediction error on the test data [6]. Also, the expectations $\mathbb{E}\{.\}$ in (24) are approximated by the sample mean over time samples. The average of the mentioned metrics and also an example average ($M = 7$) to the graph structures over thirty Monte-Carlo realizations are illustrated in Figure 1 (a:c), and (d:f), respectively. Note that although the underlying true graphs are acyclic, the plotted averaged ones can have cycles. From these results, it can be admitted that, on average, the graph recovery is superior in small values of M, i.e., $M = 2$, in case of having enough time samples, i.e., $K = 900$. Besides, the higher the number of the time samples K, the lower the prediction error err_ϵ [6]. Also, the $err_{\mathbf{c}}$ is small and robust against the changes of M if the number of time samples is not very small. i,e, $K > 300$.

B. Comparison with State-of-the-art VAR-LiNGAM [5]

In this subsection, we compare the graph recovery performance of the CGP-LiNGAM with the celebrated VAR-LiNGAM in Figure 1 (g) by varying K and M in the spans of $\{200 : 600\}$ and $\{2 : 9\}$, respectively. The polynomial coefficients $\mathbf{c}$ are generated as $2^{i+j}c_{ij} \sim \mathcal{N}(1, 0.01)$. It is well illustrated in this figure that the CGP-LiNGAM is more robust and also superior compared to the VAR-LiNGAM, especially in a small number of time samples K, i.e., $K \leq 500$, and also in high values of causal dependency M, i.e., $M \geq 5$.

C. Comparison with State-of-the-art DYNOTEARS [20]

The graph recovery results of the CGP-LiNGAM and DYNOTEARS [20] methods in different CGP order $M \in \{2, 5, 7\}$ and across time samples $K \in \{500, 1000, 2000, 5000\}$ are shown in Figure 3. It can be seen that a similar trend of improving the graph recovery with increasing the sample size K is observed in this figure for both methods; however, the DYNOTEARS

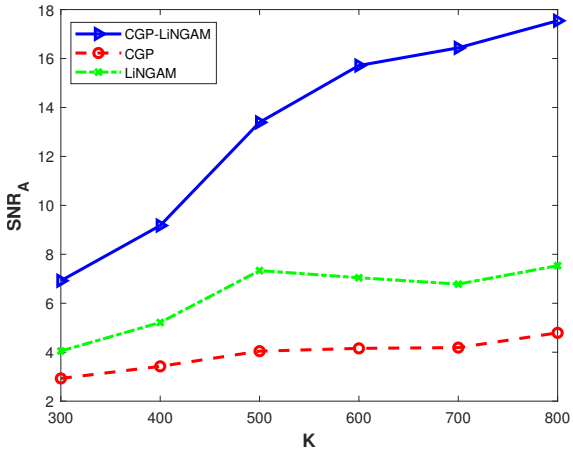

Fig. 2: Comparison of the graph recovery performance over 100 realizations for each K with the generic CGP [6] and LiNGAM [4] approaches.

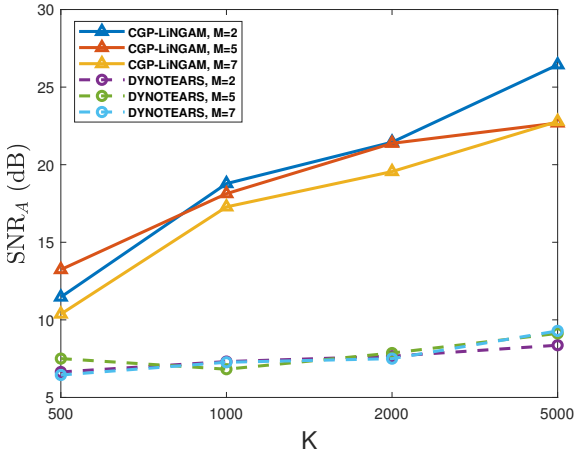

Fig. 3: Comparison of the graph recovery performance with the DYNOTEARS [20] method.

[20] has clearly failed to successfully recover the true DAGs, while the CGP-LiNGAM has estimated them with high quality. The DYNOTEARS model can be stated as [20]:

$$\mathbf{x}(k) = \mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{M} \mathbf{W}^{(i)}\mathbf{x}(k-i) + \mathbf{e}(k), \qquad (25)$$

where *only* $\mathbf{A}$ has been assumed to be an acyclic graph, and $\{\mathbf{W}^{(i)}\}_{i=1}^{M}$ are unconstrained coefficient matrices (like VAR-LiNGAM [5]) leading to an increased number of learnable parameters to $\frac{N(N-1)}{2} + MN^2$, in comparison to the proposed method which has significantly reduced $\frac{N(N-1)}{2} + \frac{M(M+3)}{2}$ learnable ones, because often $M \ll N$ [5], [6].

D. Comparison with CGP [6] and LiNGAM [4]

In this subsection, we experimentally demonstrate the effectiveness and superiority of the proposed CGP-LiNGAM model (7) over each of the CGP [6] (5) and LiNGAM [4] (7). In this way, similar to the generation scheme of Subsection IV-B and with settings $M = 2$ and $K \in \{300, 400, 500, 600, 700, 800\}$,

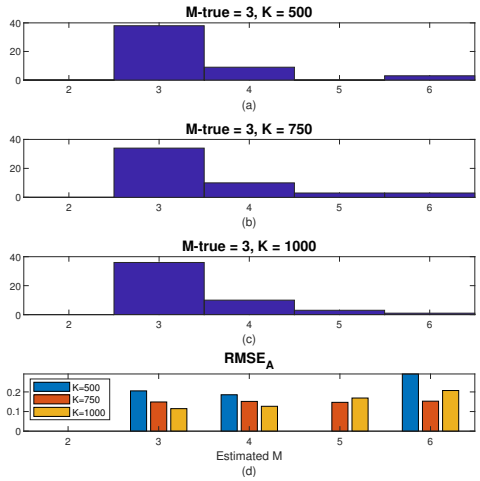

Fig. 4: (a)-(c): The histograms of the estimated M, in each $K \in \{500, 750, 1000\}$ based on minimum nAIC (26), (d): The averaged $\mathbf{RMSE_A} = \frac{\|\hat{\mathbf{A}}-\mathbf{A}\|_F}{\|\mathbf{A}\|_F}$ over realizations of each selected M.

we generate 100 realizations for each K, and the averaged DAG recovery performances are illustrated in Figure 2. Based on these results, it can be admitted that the individual models CGP (5) and LiNGAM (7) are not sufficient to fully describe the CGP-LiNGAM realizations, which shows the asymptotic efficiency of the proposed CGP-LiNGAM model (7) for accurate DAG recovery with an increase of the data samples K. Note that the CGP model is not specified for DAGs, and this is probably the main reason for lower DAG recovery performance rather than the LiNGAM approach, in which the DAGness is accounted for efficient recovery.

E. Choosing optimal CGP-order M

In most real applications, the CGP/VAR-order M is considered a fix or a prior known [6], [5]. However, due to some uncertainty conditions, e.g., existing high amount of noise, choosing it optimally can considerably boost the recovery/modeling performance [24]. In this way, to investigate the possibility of correct selection of M, we generate fifty Monte-Carlo realizations similar to Section IV-B with CGP-order $M_{true} = 3$ and the number of time samples $K \in \{500, 750, 1000\}$. Then, in each K, the normalized Akaike's Information Criterion (nAIC) (26) [24] is calculated in the span of $M \in \{2 : 6\}$ and the CGP-order M leading to minimum nAIC is considered as the estimation of M_{true}.

$$nAIC_{\hat{M}} = \log\left(\det\left(\frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{e}}(k)\hat{\mathbf{e}}(k)^{\top}\right)\right) + \frac{2n_p}{K}, \qquad (26)$$

where $det(.)$ denotes the determinant operation, $\hat{\mathbf{e}}$ is the estimated disturbances corresponding to $\hat{M}$ and the number of the estimated parameters $n_p = \frac{N(N-1)}{2} + \frac{\hat{M}(\hat{M}+3)}{2}$. Figure 4 (a)-(c) show the histograms of the estimated M, in each K. These results show that, in almost every K, M_{true} is successfully recovered. Moreover, in Figure 4 (d), the averaged

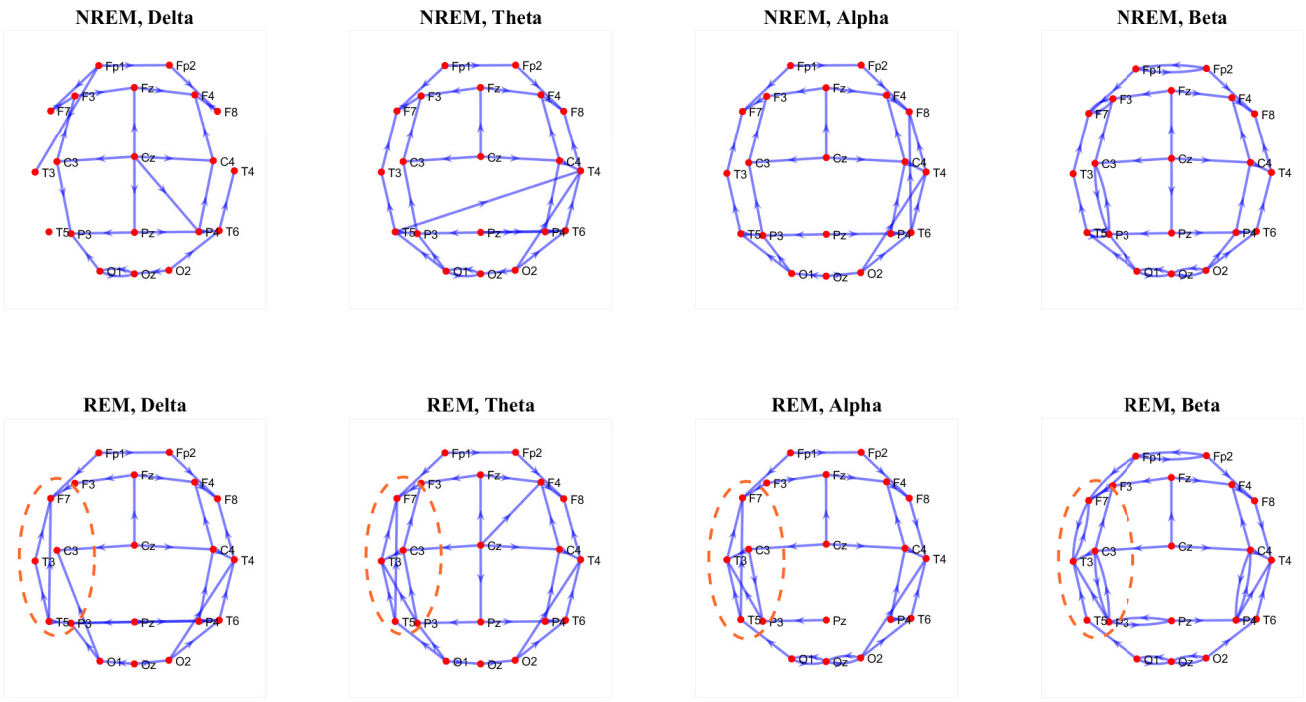

Fig. 5: The averaged effective brain connectivity corresponding to REM and Non-REM (NREM) sleep stages for each Delta, Theta, Alpha, and Beta sub-bands. Generally, it can be seen that the REM connectivity is significantly more active (having more and stronger edges), especially in the left temporal and frontal regions (specified as dashed circles in this figure), rather than the NREM ones.

$\mathbf{RMSE_A} = \frac{\|\hat{\mathbf{A}}-\mathbf{A}\|_F}{\|\mathbf{A}\|_F}$ over realizations of each selected M is plotted, which shows that, even in selected wrong M cases, i.e., $M \neq M_{opt}$, the graph recovery remains fairly robust.

F. Real Data: Learning Brain Connectivity during Sleep

To explore the usage and performance of the proposed CGP-LiNGAM method on real-world applications, we analyze its application in learning effective (directed) brain connectivity during sleep, which has been widely studied by neuroscientists for detecting a wide range of sleep disorders and measure the quality of sleep [22], [33], [9]. In this way, we have selected 200 thirty-second multi-channel sleep epochs of Subject 01 in the public and well-known MASS-SS3 dataset [34], in which each sleep epoch was scored as Wake, N1, N2, N3, or REM (sleep stages) based on the AASM [35] sleep staging standards. We considered 20 Electroencephalogram (EEG) signals (with a sampling frequency of 256 Hz) as the input data to the proposed CGP-LiNGAM model (7). The main goal is to infer the effective brain connectivity from each epoch and associated with the well-known sub-band frequencies Delta ($<$ 4 Hz), Theta (4-7 Hz), Alpha (8-15 Hz), and Beta (16-31 Hz), which is a well-studied approach for processing brain connectivity during sleep [8].

Note that after obtaining the brain connectivity for each sleep stage, the appropriate threshold to binarize the averaged graphs was determined via a statistical procedure as follows. We considered each pair of graphs for each pair of sleep stages, subtracted from each other, and then binarized this obtained subtracted graph across a span thresholds $\{0 : 0.2 : 1\}$. Afterward, we selected the final threshold, which led to the most statistical significance performed by t-test procedure. Figure 5 illustrates the averaged effective brain connectivity corresponding to REM and Non-REM (NREM) sleep stages for each Delta, Theta, Alpha, and Beta sub-bands. Firstly, from these results, it can be seen that the REM connectivity is significantly more active (having more and stronger edges), especially in the left temporal and frontal regions (specified as dashed circles in this figure), rather than the NREM ones. Secondly, in Delta and Theta bands, the directed connections from occipital to parietal and central regions have been increased for the REM stage. Regarding the Alpha and Beta sub-bands, respectively, more activity between the left parietal and frontal regions to the frontal and central ones are observed. It is worth noting that these findings are majorly supported by the previous neuroscientific studies on the estimation of effective brain connectivity during sleep [8], which shows the real-world interpretability aspect of the proposed CGP-LiNGAM approach.

V. Conclusion

In this paper, we proposed the CGP-LiNGAM method to reveal both the instantaneous and time-lagged causal relationships in time series by considering the node-specific time samples as graph signals on underlying causal DAGs. Compared to the state-of-the-art VAR-LiNGAM, our method has significantly fewer parameters, deals with only *one* underlying causal graph, and performs more robustly against a low number of time samples and a high degree of causal dependency. Our

future research direction is extending the CGP-LiNGAM to the time-varying graphs [36], [37]. Besides, the compatibility of the inferred effective brain connectivity during sleep with sleep-related neuroscientific studies illustrates the real-world applicability and interpretability of the proposed approach.

VI. Appendix (Simplifications of eq. (14))

Due to the convexity of (14) w.r.t $\tilde{\mathbf{R}}_i$ (by fixing $\{\tilde{\mathbf{R}}_{j\neq i}\}_{j=1}^{M}$) [6], and with the definitions $\mathbf{X}_m = (\mathbf{x}(m), \mathbf{x}(m+1), ..., \mathbf{x}(m+K-M-1))$ and $\tilde{\mathbf{r}}_i = vec(\tilde{\mathbf{R}}_i)$, the closed-form solutions can be obtained as:

$$\begin{aligned}
\hat{\tilde{\mathbf{R}}}_i &= \underset{\tilde{\mathbf{R}}_i}{\operatorname{argmin}} \frac{1}{2} \left\| \mathbf{X}_M - \sum_{j=1}^{M} \tilde{\mathbf{R}}_j \mathbf{X}_{M-j} \right\|_F^2 + \lambda_1 \|vec(\tilde{\mathbf{R}}_1)\|_1 \\
&+ \lambda_3 \sum_{j\neq i} \left\| [\tilde{\mathbf{R}}_i, \tilde{\mathbf{R}}_j] \right\|_F^2 \\
&= \underset{\tilde{\mathbf{R}}_i}{\operatorname{argmin}} \frac{1}{2} \left\| vec\left(\tilde{\mathbf{R}}_i \mathbf{X}_{M-i}\right) - \overbrace{vec\left(\mathbf{X}_M - \sum_{j=1\neq i}^{M} \tilde{\mathbf{R}}_j \mathbf{X}_{M-j}\right)}^{\mathbf{y}_i} \right\|_2^2 \\
&+ \lambda_1 \|vec(\tilde{\mathbf{R}}_1)\|_1 + \lambda_3 \sum_{j\neq i} \left\| vec\left(\tilde{\mathbf{R}}_i \tilde{\mathbf{R}}_j\right) - vec\left(\tilde{\mathbf{R}}_j \tilde{\mathbf{R}}_i\right) \right\|_2^2.
\end{aligned} \tag{27}$$

Then, using the relation $vec(\mathcal{ABC}) = (\mathcal{C}^\top \otimes \mathcal{A}) vec(\mathcal{B})$ [28]:

$$\begin{aligned}
\hat{\tilde{\mathbf{r}}}_i &= \underset{\tilde{\mathbf{r}}_i}{\operatorname{argmin}} \frac{1}{2} \left\| \overbrace{\left(\mathbf{X}_{M-i}^\top \otimes \mathbf{I}\right)}^{\mathbf{B}_i} \tilde{\mathbf{r}}_i - \mathbf{y}_i \right\|_2^2 + \lambda_1 \|\tilde{\mathbf{r}}_1\|_1 \\
&+ \lambda_3 \sum_{j\neq i} \left\| \overbrace{\left(\tilde{\mathbf{R}}_j^\top \otimes \mathbf{I} - \mathbf{I}^\top \otimes \tilde{\mathbf{R}}_j\right)}^{\mathbf{\Phi}_j} \tilde{\mathbf{r}}_i \right\|_2^2 \\
&= \underset{\tilde{\mathbf{r}}_i}{\operatorname{argmin}} \|\mathbf{\Psi}_i \tilde{\mathbf{r}}_i - \tilde{\mathbf{y}}_i\|_2^2 + \lambda_1 \|\tilde{\mathbf{r}}_1\|_1.
\end{aligned} \tag{28}$$

References

[1] Judea Pearl. *Causality*. Cambridge university press, 2009.

[2] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[3] Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

[4] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[5] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

[6] Jonathan Mei and José MF Moura. Signal processing on graphs: Causal modeling of unstructured data. *IEEE Transactions on Signal Processing*, 65(8):2077–2092, 2016.

[7] Junzhong Ji, Aixiao Zou, Jinduo Liu, Cuicui Yang, Xiaodan Zhang, and Yongduan Song. A survey on brain effective connectivity network learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[8] Aminollah Glorou, Ali Sheikhani, Ali Motie Nasrabadi, and Mohammad Reza Saebipour. Detecting slow wave sleep and rapid eye movement stage using cortical effective connectivity. *Turkish Journal of Electrical Engineering and Computer Sciences*, 26(6):2779–2791, 2018.

[9] Basak Alper, Benjamin Bach, Nathalie Henry Riche, Tobias Isenberg, and Jean-Daniel Fekete. Weighted graph comparison techniques for brain connectivity analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 483–492, 2013.

[10] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

[11] Xiaowen Dong, Dorina Thanou, Laura Toni, Michael Bronstein, and Pascal Frossard. Graph signal processing for machine learning: A review and new perspectives. *IEEE Signal processing magazine*, 37(6):117–127, 2020.

[12] Antonio Ortega. *Introduction to graph signal processing*. Cambridge University Press, 2022.

[13] Xianming Liu, Gene Cheung, Xiangyang Ji, Debin Zhao, and Wen Gao. Graph-based joint dequantization and contrast enhancement of poorly lit jpeg images. *IEEE Transactions on Image Processing*, 28(3):1205–1219, 2018.

[14] Gene Cheung, Enrico Magli, Yuichi Tanaka, and Michael K Ng. Graph spectral image processing. *Proceedings of the IEEE*, 106(5):907–930, 2018.

[15] Chenggang Yan, Zhisheng Li, Yongbing Zhang, Yutao Liu, Xiangyang Ji, and Yongdong Zhang. Depth image denoising using nuclear norm and learning graph model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(4):1–17, 2020.

[16] Anindya Mondal, Jhony H Giraldo, Thierry Bouwmans, Ananda S Chowdhury, et al. Moving object detection for event-based vision using graph spectral clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 876–884, 2021.

[17] Jhony H Giraldo, Sajid Javed, Maryam Sultana, Soon Ki Jung, and Thierry Bouwmans. The emerging field of graph signal processing for moving object segmentation. In *International Workshop on Frontiers of Computer Vision*, pages 31–45. Springer, 2021.

[18] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[19] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

[20] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

[21] Weiyu Huang, Thomas AW Bolton, John D Medaglia, Danielle S Bassett, Alejandro Ribeiro, and Dimitri Van De Ville. A graph signal processing perspective on functional brain imaging. *Proceedings of the IEEE*, 106(5):868–885, 2018.

[22] Thien Nguyen, Olajide Babawale, Tae Kim, Hang Joon Jo, Hanli Liu, and Jae Gwan Kim. Exploring brain functional connectivity in rest and sleep states: a fnirs study. *Scientific reports*, 8(1):16144, 2018.

[23] Antonio G Marques, Santiago Segarra, Geert Leus, and Alejandro Ribeiro. Stationary graph processes and spectral estimation. *IEEE Transactions on Signal Processing*, 65(22):5911–5926, 2017.

[24] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[25] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[26] Andrew Bolstad, Barry D Van Veen, and Robert Nowak. Causal network inference via group sparse regularization. *IEEE transactions on signal processing*, 59(6):2628–2641, 2011.

[27] Ferenc Szidarovszky et al. *Introduction to Matrix Theory: with applications to business and economics*. World Scientific, 2002.

[28] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[29] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[30] Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

[31] Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2008.

[32] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

[33] Takamitsu Watanabe, Shigeyuki Kan, Takahiko Koike, Masaya Misaki, Seiki Konishi, Satoru Miyauchi, Yasushi Miyahsita, and Naoki Masuda. Network-dependent modulation of brain activity during sleep. *NeuroImage*, 98:1–10, 2014.

[34] Christian O'reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of sleep research*, 23(6):628–635, 2014.

[35] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V Vaughn, et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176:2012, 2012.

[36] Vassilis Kalofolias, Andreas Loukas, Dorina Thanou, and Pascal Frossard. Learning time varying graphs. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2826–2830. Ieee, 2017.

[37] Jhony H Giraldo, Arif Mahmood, Belmar Garcia-Garcia, Dorina Thanou, and Thierry Bouwmans. Reconstruction of time-varying graph signals via sobolev smoothness. *IEEE Transactions on Signal and Information Processing over Networks*, 8:201–214, 2022.