MultiFun-DAG: Multivariate Functional Directed Acyclic Graph

Tian Lan
Department of Industrial Engineering, Tsinghua University
Ziyue Li
Information Systems Department, University of Cologne
Junpeng Lin
Department of Industrial Engineering, Tsinghua University
Zhishuai Li
Sensetime
Lei Bai
Shanghai AI Laboratory
Man Li
Department of Industrial Engineering and Decision Analytics,
The Hong Kong University of Science and Technology
Fugee Tsung
Department of Industrial Engineering and Decision Analytics,
The Hong Kong University of Science and Technology
Rui Zhao
Sensetime
Chen Zhang
Department of Industrial Engineering, Tsinghua University

April 23, 2024

Abstract

Directed Acyclic Graphical (DAG) models efficiently formulate causal relationships in complex systems. Traditional DAGs assume nodes to be scalar variables, characterizing complex systems under a facile and oversimplified form. This paper considers that nodes can be multivariate functional data and thus proposes a multivariate functional DAG (MultiFun-DAG). It constructs a hidden bilinear multivariate function-to-function regression to describe the causal relationships between different nodes. Then an Expectation-Maximum algorithm is used to learn the graph structure as a score-based algorithm with acyclic constraints. Theoretical properties are diligently derived. Prudent numerical studies and a case study from urban traffic congestion analysis are conducted to show MultiFun-DAG's effectiveness.

Keywords: Causal Structure Learning, Functional Data, Directed Acyclic Graph

1 Introduction

Directed acyclic graph (DAG), a.k.a., Bayesian network, is a probabilistic graphical model that represents a set of variables and their causal relationships. In a DAG, each node corresponds to a random variable, and each directed acyclic edge represents a causal dependence relationship between the two variables, i.e., a parent node and a descendant node. The distribution of each variable can be written as a conditional probability distribution given its parent nodes and is independent from other nodes. DAG has been widely used to offer vital insights for causal relationship discovery in biological (Aguilera et al., 2011), physical (Velikova et al., 2014), social systems (Ruz et al., 2020), etc.

Previous work has thoroughly studied DAG with each node as a scalar variable (Heckerman, 2008). However, it is common to come across systems where the variables have a *functional* form, as shown in Fig. 1. (b). Functional data is formally defined as the data with each sample in the form of random curves or functions over a continuum, such as time or space (Qiao et al., 2019), which is commonly observed in complex systems such as medical science (Chen et al., 2018), physiology (Li and Solea, 2018), and climate (Fraiman et al., 2014). For example, in urban transportation, sensors collect the real-time signals of the traffic elements, such as traffic volume, vehicle speed, lane saturation, cycle length of traffic lights, and weather, which are all functional data and can be combined into a multivariate form. By modeling these traffic variables as different nodes in a DAG to learn their causal relationships, root causes for traffic congestion can be identified, and then corresponding actions can be taken (Lan et al., 2023).

We consider the DAG in which each node can be multivariate functional data, as it can describe the practical systems more pertinently than the scalar-based ones. Such a DAG has three critical properties: (1) **Infinite dimensionality**: Functional data are naturally

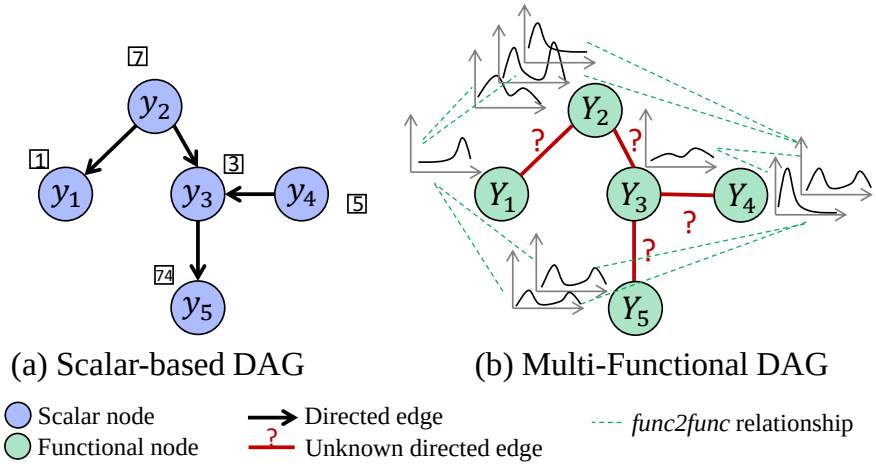

Figure 1: Scalar-based DAG v.s. Multi-Functional DAG. Each node is a scalar or functional variable, and the directed edge is the causal dependence. MultiFun-DAG learns the unknown causal edge (solid) via formulating the *func2func* relationships (dotted).

infinite-dimensional, and in theory, can have infinitely many points; Though in practice functional data is usually discretized or approximated to a finite number of observation points. However, the theoretical foundation is that the true underlying functional observation is of infinite dimensionality. (2) **Data heterogeneity**: Functions of different nodes can be heterogeneous, such as containing various numbers of functions and coming from different spaces. (3) **Inter-causation**: Functions of different nodes could be inter-correlated in different ways, i.e., different functions of one node can have different causal effects on another function of another node.

As a result, traditional scalar-based DAGs cannot be easily extended to our case.

This paper aims to build a **Multi-Fun**ctional **DAG** (MultiFun-DAG) to learn the valuable causal dependence structure among different multi-functional nodes. The task is unfolded by three concrete questions: (1) how to preserve the information and describe

causal dependence relationships for infinite functions? (2) how to model and fuse the causal dependence relationships between multiple functions in any two nodes and build an edge between them? (3) how to conduct structural learning and parameter learning for these edges?

To address these challenges, we are the first to propose a novel DAG to learn the causal structure with multivariate functional data, with the following major contributions:

- We model the causal dependence relationships between nodes with multiple functions via hidden bilinear *function-to-function (func2func) regression* with *low-rank decomposition*.
- We propose an Expectation-Maximization (EM) algorithm in the score-based structural learning framework to learn the DAG structure with acyclic constraint and group lasso penalty.
- We derive the theoretical properties of the model, including its identifiability and asymptotic error bound of the EM algorithm, and the asymptotic oracle property of our structure learning algorithm.

2 Related Work

2.1 DAG structural learning methods

Methods for DAG learning can be categorized into combinatorial learning and continuous learning algorithms.

Combinatorial learning algorithms solve a combinatorial optimization problem to find whether an edge exists between any two nodes. This type of method can be further divided into constraint-based and score-based algorithms.

Constraint-based methods, such as PC (Spirtes et al., 2000), rankPC (Harris and Drton, 2013), and fast causal inference (Spirtes et al., 2000), learn the edges by conditional independence tests. However, they are built upon that the independence tests should accurately reflect the independence model, which is generally difficult to be satisfied in reality. As a result, these methods suffer from error propagation, where a minor error in the early phase can result in a very different DAG.

The score-based methods instead construct a score function to evaluate DAG structures and select the graph with the highest score. Some commonly used score functions include the likelihood function, mean square fitting error, etc. Some further regularization items on edges are also added in the score to learn a sparse graph (Chickering, 2002; Nandy et al., 2018). Then greedy searches are implemented to find the graph with the highest score. However, one drawback of the combinatorial score-based method is the nonconvexity of the combinatorial problem. The acyclicity constraint means that the solution space stretches along all topological orderings that have $d!$ permutations in a graph with d nodes, rendering DAG learning an NP-hard problem.

Continuous learning algorithms formulate the acyclic constraint into an algebraic form and convert the structure learning problem into a purely continuous optimization problem to save computation cost. In particular, Zheng et al. (2018) proposed NoTears, which formulates an algebraic form as $h(W) = \mathrm{tr}(\exp(W \circ W)) - d = 0$, where W is the adjacency weight matrix, $\mathrm{tr}(\cdot)$ is the trace, and $\circ$ is Hadamard product. This idea was popularly borrowed in many preceding works. For example, Zheng et al. (2020) develops a nonparametric DAG based on NoTears Bhattacharya et al. (2021) considers both directed and undirected edges based on NoTears. Besides, Ng et al. (2020) also proposes a soft constraint for acyclicity. However, the NoTears-based methods only offer solutions for

scalar-variable nodes. The more realistic problem where nodes contain heterogeneous multi-functional data has never been addressed so far.

2.2 Functional graphical models

Functional graphical models (FGMs), as an extension of traditional graphical models, describe the probabilistic dependence between nodes with functional data and could potentially offer solutions for functional DAG learning. According to the direction of the edges, FGMs can be divided into undirected FGMs and directed FGMs.

The undirected FGMs focus on estimating the correlation dependence structure between different nodes. In particular, Qiao et al. (2019) proposes a functional graphical Lasso model to describe the sparse correlation dependence structure of different functional nodes. As an extension, Qiao et al. (2020) proposes a doubly FGM to capture the evolving conditional dependence among functions. Later more FGMs were proposed, such as using nonparametric additive conditional independence model (Li and Solea, 2018), assuming the dependence to be partially separable (Zapata et al., 2022), or heterogeneous (Wu et al., 2022), etc. However, undirected FGMs only capture the correlations, instead of causation, of nodes.

For directed FGMs focusing on the causal relationship of nodes, the current research is scarce. Sun et al. (2017) proposes a DAG that considers both scalar and functional nodes. Yet it assumes the DAG structure is known in advance. Gómez et al. (2020) considers DAG with each node as a univariate function. However, it still assumes the topological ordering of nodes should be known in advance by domain knowledge, and transforms the structural learning problem into a parameter selection problem, i.e., selecting the parent node from the candidate parent set. Furthermore, Gómez et al. (2020) is a two-step framework by first

adopting functional principal component analysis (FPCA) to extract features for each node separately, and then using the FPCA scores to model the causal effects. However, since its FPCA totally ignores the causal relationships between different nodes, the extracted PCs may not represent the most useful information in the whole network. Then the causal effects estimated based on these PCs may be misleading and lead to higher estimation errors.

3 Proposed Model

Suppose that a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents a DAG with a vertex set $\mathcal{V} \in \mathbb{R}^P$ and an edge set $\mathcal{E} \in \mathbb{R}^{P\times P}$, with P denoted as the total number of nodes. A tuple $(j, j') \in \mathcal{E}$ represents a directed edge leading from node j to node j', i.e., $j \to j'$. Here we assume the node j has L_j functional variables, with $Y_{jl}(t), t \in \Gamma$ denoted as its l-th function, for $l = 1, 2, ..., L_j$. Here without loss of generality, we assume $\Gamma = [0, 1]$ is a compact time interval. Suppose we have N identically and independently distributed samples. The n-th sample, $n = 1, \ldots, N$, is formulated as $\mathbf{Y}^{(n)}(t) = (\mathbf{Y}_1^{(n)}(t), \mathbf{Y}_2^{(n)}(t), ..., \mathbf{Y}_P^{(n)}(t))^T$, with $\mathbf{Y}_j^{(n)}(t) = (Y_{j1}^{(n)}(t), Y_{j2}^{(n)}(t), ..., Y_{jL_j}^{(n)}(t))$. Therefore, $\mathbf{Y}^{(n)}(t)$ represents $L = \sum_j L_j$ functions of all the nodes, which is a vector. Our MultiFun-DAG aims to learn the causal relations between different nodes, i.e., the edge set $\mathcal{E}$, shown as red lines in Fig. 1.

To achieve it, in Section 3.1, we first assume that the causal structure $\mathcal{E}$ is known, and construct a hidden bilinear *func2func* regression, to learn the conditional dependence from the function l of node j to the function l' of node j', shown as the green dotted edge in Fig. 1. In Section 3.2, we show that causal structure is non-identifiable under maximum likelihood estimation. Therefore, we introduce a restriction for DAG structure and its necessity. In Section 3.3, we combine the restriction in Section 3.2 and propose an EM

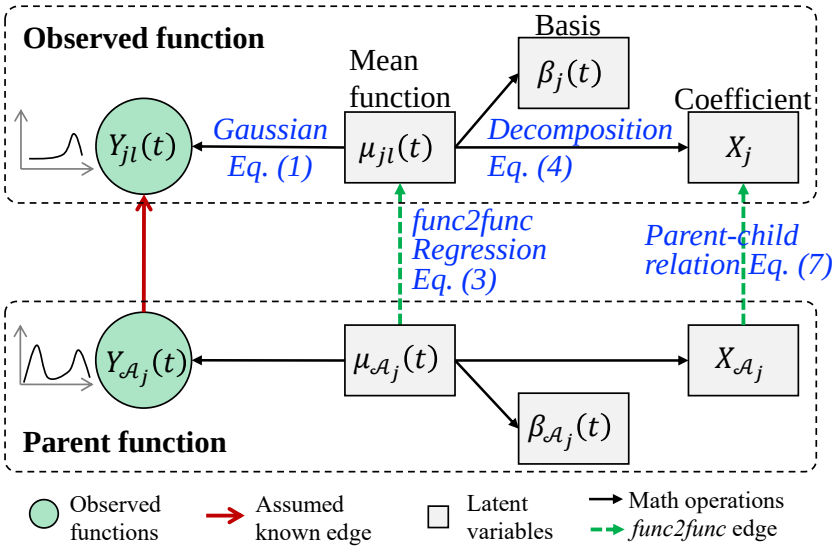

Figure 2: The Illustration of MultiFun-DAG

algorithm for learning the causal structure of MultiFun-DAG.

3.1 Multi-functional DAG with known structure

We first give an overview of our MultiFun-DAG in Fig. 2. Our function $Y_{jl}(t)$ follows Gaussian distribution in Eq. (1) with mean function $\mu_{jl}(t)$. The mean functions follow *func2func* regression in Eq. (3) with their parents in DAG. To preserve the information for infinite functional variables, we decompose the mean function into a basis set with coefficients in Eq. (4). Then we conduct a bilinear regression for the coefficients to describe the linear causality of different nodes via Eq. (7). The joint likelihood of coefficients of all the nodes can be represented using a linear Structural Equation Model (SEM) (Eq. (8)).

In this paper, we focus on Gaussian distributed function:

$$Y_{jl}^{(n)}(t) \sim \mathcal{N}(\mu_{jl}^{(n)}(t), R_{jl}(\cdot,\cdot)), \tag{1}$$

where $\mu_{jl}^{(n)}(t)$ is the mean function and $R_{jl}(\cdot,\cdot)$ is the covariance function of $Y_{jl}^{(n)}$. We

assume that $R_{jl}(t,t') = r_{jl}^2 \mathbb{I}(t = t')$, where r_{jl}^2 is the scale of variance. The parent set of node j is denoted as $\mathcal{A}_j = \{j' | j' \in \mathcal{V}, j' \neq j, (j', j) \in \mathcal{E}\}$. We assume that the joint distribution of $\mu_{jl}^{(n)}(t)$ of all the nodes can be written as the production of the conditional distribution of each node, i.e.,

$$p(\mu_{11}^{(n)}(t), \ldots, \mu_{PL_P}^{(n)}(t)) = \prod_{j=1}^{P} \prod_{l=1}^{L_j} p(\mu_{jl}^{(n)}(t) | \mathcal{A}_j). \tag{2}$$

We focus on linear conditional dependence relationship for $p(\mu_{jl}^{(n)}(t)|\mathcal{A}_j)$, which is formulated as below:

$$\mu_{jl}^{(n)}(t) = \sum_{j' \in \mathcal{A}_p} \sum_{l'=1}^{L_{j'}} \int_0^1 \gamma_{j'jl'l}(t,s) \mu_{j'l'}^{(n)}(s) \mathrm{d}s + \varepsilon_{jl}^{(n)}(t), \tag{3}$$

where $\varepsilon_{jl}^{(n)}(t)$ is the noise function. $\gamma_{j'jl'l}(t,s)$ is the coefficient function for $(j', j) \in \mathcal{E}$, $l = 1, 2, ..., L_j$ and $l' = 1, 2, ..., L_{j'}$, which describes the contribution of the l'-th function of node j' to the l-th function of node j. We represent $\gamma_{j'jl'l}(t,s)$ and $\mu_{jl}(t)$ as follows:

For $\mu_{jl}(t)$: Given they are in infinite dimensions and hard to be estimated directly, it is common to decompose them into a well-defined continuous space for feature extraction:

$$\mu_{jl}^{(n)}(t) = \sum_{k=1}^{K_j} x_{jlk}^{(n)} \beta_{jk}(t), \tag{4}$$

where $\mathbf{B}_j(t) = (\beta_{j1}(t), \beta_{j2}(t), ..., \beta_{jK_j}(t))^T$ is an orthonormal functional basis set for node j, with $\int \beta_{jk}(t)^2 \mathrm{d}t = 1, k = 1, \ldots, K_P$ and $\int \beta_{jk}(t)\beta_{jk'}(t)\mathrm{d}t = 0,\ k \neq k'$. $x_{jlk}^{(n)}$ is the corresponding coefficient.

For $\gamma_{j'jl'l}(t,s)$: , we describe $\gamma_{j'jl'l}(t,s)$ using the corresponding basis sets in a bilinear way (Hoff, 2015) as:

$$\gamma_{j'jl'l}(t,s) = \sum_{k=1}^{K_j} \sum_{k'=1}^{K_{j'}} c_{j'jk'k} \cdot c_{j'jl'l} \beta_{j'k'}(s) \beta_{jk}(t). \tag{5}$$

$c_{j'jk'k}$ represents the influence caused by the basis pair: $\beta_{j'k'}(s)$ on $\beta_{jk}(t)$. $c_{j'jl'l}$ represents the influence caused by the function pair: function l' of node j' on function l of node j.

This decomposition describes the regression coefficient function from two aspects, i.e., (1) the basis set of a node and (2) the variables of a node, separately. Besides, it also improves estimation stability.

By plugging the representation of Eq. (4) and (5) into Eq. (3), for function l in node j, we could obtain:

$$\sum_{k=1}^{K_j} x_{jlk}^{(n)} \beta_{jk}(t) = \sum_{j' \in \mathcal{A}_j} \sum_{k=1}^{K_j} \sum_{l'=1}^{L_{j'}} \sum_{k'=1}^{K_{j'}} \int_0^1 c_{j'jk'k} \cdot c_{j'jl'l} x_{j'l'k'}^{(n)} \beta_{jk}(t) \beta_{j'k'}^2(s) \mathrm{d}s + \varepsilon_{jl}(t). \tag{6}$$

By integrating this equation over s, and combining all the parameters $x_{jlk}^{(n)}$ into a vector, i.e., $\mathbf{x}_j^{(n)} = \mathrm{vec}(x_{jlk}^{(n)}) \in \mathbb{R}^{L_j K_j}$, where $[\mathbf{x}_j^{(n)}]_i$ represents the $[(i-1) \bmod K_j] + 1$ coefficient of the function $\lfloor (i-1)/K_j \rfloor + 1$ in node j, Eq. (6) can be re-written as:

$$\mathbf{x}_j^{(n)} = \sum_{j' \in \mathcal{A}_j} (\mathbf{C}_{j'j}^L \otimes \mathbf{C}_{j'j}^K)^T \mathbf{x}_{j'}^{(n)} + \boldsymbol{\xi}_j^{(n)}. \tag{7}$$

Here $\mathbf{C}_{j'j}^L \in \mathbb{R}^{L_{j'} \times L_j}$ with $[\mathbf{C}_{j'j}^L]_{l'l} = c_{j'jl'l}, \mathbf{C}_{j'j}^K \in \mathbb{R}^{K_{j'} \times K_j}$ with $[\mathbf{C}_{j'j}^K]_{k'k} = c_{j'jk'k}$. $\otimes$ is the Kronecker product. $\boldsymbol{\xi}_j \in \mathbb{R}^{L_j K_j}$ is the noise of $\mathbf{x}_j$, where $[\boldsymbol{\xi}_j]_{(l-1)K_j+1}$ to $[\boldsymbol{\xi}_j]_{lK_j}$ are the projection of $\varepsilon_{jl}^{(n)}(t)$ on its corresponding basis set for $j = 1, \ldots, P, l = 1, \ldots, L_j$. Here we assume $\boldsymbol{\xi}_j^{(n)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_j)$ with $\boldsymbol{\Omega}_j \in \mathbb{R}^{L_j K_j \times L_j K_j}$. For brevity, we simply assume $\boldsymbol{\Omega}_j$ has a diagonal form, i.e., $\boldsymbol{\Omega}_j = \mathrm{diag}(\boldsymbol{\omega}_j^2)$.

Lastly, we use a linear SEM to interpret our MultiFun-DAG. We denote $\mathbf{C} \in \mathbb{R}^{M \times M}$ with its (j, j') block as $\mathbf{C}_{(j',j)} = \mathbf{C}_{j'j}$, $\mathbf{C}_{j'j} = \mathbf{C}_{j'j}^L \otimes \mathbf{C}_{j'j}^K$ if $(j, j') \in \mathcal{E}$, otherwise we have $\mathbf{C}_{(j',j)} = \mathbf{0}_{L_{j'} K_{j'} \times L_j K_j}$. Then for $\mathbf{x}^{(n)} = [\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_P^{(n)}] \in \mathbb{R}^M$, a linear SEM interpretation is:

$$\mathbf{x}^{(n)} = \mathbf{C}^T \mathbf{x}^{(n)} + \boldsymbol{\xi}^{(n)}. \tag{8}$$

Here $M = \sum_{j=1}^P L_j K_j$, $\boldsymbol{\xi}^{(n)} = [\boldsymbol{\xi}_1^{(n)}, \ldots, \boldsymbol{\xi}_P^{(n)}] \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ is the noise vector. $\boldsymbol{\Omega} = \mathrm{diag}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_P)$.

In reality, $Y_{jl}(t)$ can only be measured at certain discrete observation points. In this paper, without loss of generality, we assume that, for all the nodes, the sampling points are equally spaced as $t_1, \ldots, t_T$. Then we define $\mathbf{X} = [\mathbf{x}^{(1)T}, \ldots, \mathbf{x}^{(N)T}]^T \in \mathbb{R}^{N\times M}$, $\mathbf{Y}_{jl}^{(n)} = [Y_{jl}^{(n)}(t_1), \ldots, Y_{jl}^{(n)}(t_T)]$. By abusing the notation $\mathbf{Y}^{(n)} = [\mathbf{Y}_{jl}^{(n)}, j = 1, \ldots, P, l = 1, \ldots, L_j]$ and $\mathcal{Y} = [\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(N)}]$ for convenience, we can write the joint likelihood of the generative model as:

$$f(\mathbf{X}, \mathcal{Y}) = \prod_{i=1}^{N} p(\mathbf{x}^{(n)}) p(\mathbf{Y}^{(n)}|\mathbf{x}^{(n)}), \tag{9}$$

where $p(\mathbf{x}^{(n)})$ and $p(\mathbf{Y}^{(n)}|\mathbf{x}^{(n)})$ are computed by Eqs. (1), (4) and (8). It is to be noted that our model can also be applicable to functional nodes measured at distinct observation points with different lengths, with trivial notation modifications.

3.2 Non-identifiability and equivalence class

In reality, the graph structure is unknown and to be estimated. This can be transferred to infer whether the weight $\mathbf{C}_{j'j}^{L}$ and $\mathbf{C}_{j'j}^{K}$ equals $\mathbf{0}$ for certain blocks. In particular, the parameters to be estimated in our model includes 1) the weights $\mathbf{C}_{j'j}^{L}$ and $\mathbf{C}_{j'j}^{K}$ for nodes $j, j' = 1, \ldots, P$; 2) the variance of functional noise, denoted as $\mathbf{r} = [r_{11}^2, r_{12}^2, ..., r_{PL_P}^2] \in \mathbb{R}^M$; 3) the variance of $\mathbf{x}_j$, i.e., $\mathbf{\Omega}_1, \mathbf{\Omega}_2, ..., \mathbf{\Omega}_P$, denoted as $\mathbf{\Omega}_{[1:P]}$; 4) the basis functions $\mathbf{B}_j(t) = [\beta_{j1}(t), \beta_{j2}(t), ..., \beta_{jK_j}(t)]^T$ for node $j = 1, \ldots, P$. It is to be noted that in reality, we only need to estimate $\mathbf{B}_j = [\mathbf{B}_j(t_1)^T, \ldots, \mathbf{B}_j(t_T)^T]^T$, denoted as $\mathbf{B} = [\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_P]$. For other observation points, we can adopt Kernel smoothing to estimate them easily.

The parameters $\Theta = (\mathbf{C}, \mathbf{B}, \mathbf{r}, \mathbf{\Omega}_{[1:P]})$ are statistically nonidentifiable without further constraints. Based on our model structure, the marginal distribution of $\mathbf{Y}$ follows a Gaussian distribution with mean $\mathbf{0}$ and covariance function $\mathbf{\Sigma}_{\mathbf{Y}}(\Theta)$, which is determined by the

parameters Θ, i.e.,

$$\mathbf{\Sigma_Y}(\Theta)_{jl,j'l'} = \begin{cases} \mathbf{B}_j[(\mathbf{I}-\mathbf{C})^{-T}\mathbf{\Omega}(\mathbf{I}-\mathbf{C})^{-1}]_{jl,jl}\mathbf{B}_j^T + r_{jl}^2\mathbf{I}_T & (j,l)=(j',l') \\ \mathbf{B}_j[(\mathbf{I}-\mathbf{C})^{-T}\mathbf{\Omega}(\mathbf{I}-\mathbf{C})^{-1}]_{jl,j'l'}\mathbf{B}_{j'}^T & o.w. \end{cases}. \tag{10}$$

We aim to estimate the model parameters Θ based on the information from the observed covariance matrix $\mathbf{\Sigma_Y}$. However, it turns out that the mapping from Θ to $\mathbf{\Sigma_Y}(\Theta)$ is not one-to-one, i.e., one $\mathbf{\Sigma_Y}$ can correspond to multiple sets of model parameters Θ. Denote the true covariance matrix as $\mathbf{\Sigma_Y^*} = \mathbf{\Sigma_Y}(\Theta^*)$, where Θ^* is the true underlying parameters. We define the set of all Θ whose $\mathbf{\Sigma_Y}(\Theta)$ equals $\mathbf{\Sigma_Y^*}$ as the equivalence class $\mathfrak{D}$ corresponding to $\mathbf{\Sigma_Y^*}$, i.e.,

$$\mathfrak{D} = \{\Theta | \mathbf{\Sigma_Y^*} = \mathbf{\Sigma_Y}(\Theta)\}.$$

Without additional restrictions, we can only find one $\Theta \in \mathfrak{D}$ based on the observation data. However, infinite combinations of parameters exist in the equivalence class and cannot provide us with useful information regarding the causal structure. The most common solution for Gaussian noise is to assume Condition 1, which can be viewed as an extension of the equal variance condition in Van de Geer and Bühlmann (2013).

Condition 1. *In the true DAG, all latent variables have equal variance, i.e.,* $\mathbf{\Omega} = \omega_0^2\mathbf{I}$

It is a common condition for ensuring the identifiability of a linear structural causal model with Gaussian noise. With it, all the graphs with $\Theta \in \mathfrak{D}$ will have the same causal structure.

3.3 Regularized EM estimation

Under Condition 1, we rewrite the parameter set as $\Theta = \{\mathbf{C}, \mathbf{B}, \mathbf{r}, \omega_0^2\}$. Since the coefficients $\mathbf{X}$ are unknown, we estimate $\mathbf{x}^{(n)}, n = 1, \ldots, N$ by treating them as latent variables and

using a regularized EM algorithm for estimation (Yi and Caramanis, 2015). The regularized EM algorithm consists of an Expectation-step and a regularized Maximization-step. In each iteration, the operator $\mathcal{M}_n$ of the regularized EM is denoted as follows:

$$\begin{aligned} \mathcal{M}_n(\Theta') &= \arg\max_{\Theta} Q_n(\Theta;\Theta') - \lambda\mathcal{R}(\mathbf{C}) \\ s.t. \qquad & \mathcal{G} \text{ is a DAG}, \end{aligned} \tag{11}$$

where

$$Q_n(\Theta;\Theta') = \mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta'} \log f(\mathbf{X},\mathcal{Y};\Theta) = \int \log f(\mathbf{X},\mathcal{Y};\Theta') p(\mathbf{X}|\mathcal{Y};\Theta') \mathrm{d}\mathbf{X}, \tag{12}$$

$$\begin{aligned} \log f(\mathbf{X},\mathcal{Y};\Theta) = -\frac{1}{2}\Bigg(\sum_{n=1}^{N} \Big(\sum_{j=1}^{P}\sum_{l=1}^{L_j} (\mathbf{Y}_{jl}^{(n)} - \mathbf{B}_j\mathbf{x}_{jl}^{(n)})^T r_{jl}^{-2} (\mathbf{Y}_{jl}^{(n)} - \mathbf{B}_j\mathbf{x}_{jl}^{(n)}) \\ + \sum_{j=1}^{P} (\mathbf{x}_j^{(n)} - \mathbf{x}^{(n)}\mathbf{C}_j)^T \omega_0^{-2} (\mathbf{x}_j^{(n)} - \mathbf{x}^{(n)}\mathbf{C}_j) + \sum_{j=1}^{P}\sum_{l=1}^{L_j} T \log r_{jl}^2 \\ + M \log \omega_0^2 \Big) \Bigg) + constants, \end{aligned} \tag{13}$$

and $\mathcal{R}(\mathbf{C})$ is the sparse penalty, to penalize the model complexity.

To represent the DAG constraint in Eq. (11) to a mathematical form, we define the adjacency matrix $\mathbf{W} \in \mathbb{R}^{P\times P}$ corresponding to the edge set $\mathcal{E}$ for the DAG $\mathcal{G}$. Consider $\mathbf{W}$ as a measure of causal effects and it fuses the information of $\mathbf{C}_{ij}$ in a scalar. We have:

$$[\mathbf{W}]_{ij} \neq 0 \Leftrightarrow \mathbf{C}_{ij} \neq \mathbf{0}_{L_iK_i \times L_jK_j}. \tag{14}$$

Then in this work, we give an intuitive and valid definition for $\mathbf{W}$ as

$$[\mathbf{W}]_{ij} \doteq \|\mathbf{C}_{ij}\|_F. \tag{15}$$

Consequently, to ensure $\mathbf{C}$ is a DAG, we adopt Notears constraints (Zheng et al., 2018) for the adjacency matrix $\mathbf{W}$ that $h(\mathbf{W}) := \mathrm{tr}(\exp(\mathbf{W} \circ \mathbf{W})) - P$ and we have:

$$h(\mathbf{W}) = 0 \Leftrightarrow \mathcal{G} \text{ is a DAG}. \tag{16}$$

Finally, for a large graph, it is usually assumed the edges are sparse, and penalize the l_1 norm of $\mathbf{W}$. Therefore, we set $\mathcal{R}(\mathbf{C}) = \|\mathbf{C}\|_{l_1/F} = \sum_{i=1}^{P}\sum_{j=1}^{P}\|\mathbf{C}_{ij}\|_F$, and λ adjusts the strength of the penalty.

Expectation-step is to calculate $Q_n(\Theta;\Theta')$. It can be derived by calculating the posterior likelihood $p(\mathbf{X}|\mathcal{Y};\Theta')$, which can be estimated in a forward and backward way.

Proposition 1. *For any parameter set Θ', the posterior distribution can be decomposed as $p(\mathbf{X}|\mathcal{Y};\Theta') = \prod_{i=1}^{N} p(\mathbf{x}^{(n)}|\mathbf{Y}^{(n)};\Theta')$. $p(\mathbf{x}^{(n)}|\mathbf{Y}^{(n)};\Theta')$ follows a multivariate normal distribution $\mathcal{N}(\hat{\mathbf{u}}^{(n)}_{\Theta',\mathbf{Y}^{(n)}}, \hat{\mathbf{\Sigma}}_{\Theta'})$ with mean $\hat{\mathbf{u}}^{(n)}_{\Theta',\mathbf{Y}^{(n)}} \in \mathbb{R}^{1\times P}$ and variance $\hat{\mathbf{\Sigma}}_{\Theta'} \in \mathbb{R}^{P\times P}$, where $\hat{\mathbf{u}}_{\Theta,\mathbf{Y}}$ is a linear combination of $\mathbf{Y}$ depending on Θ while $\hat{\mathbf{\Sigma}}_{\Theta}$ only depends on Θ.*

Proof. Proposition 1 is straightforward by following the procedure in Appx. B.3. □

Maximization-step is to solve the maximization problem of Eq. (11) based on the calculated $Q_n(\Theta,\Theta')$ in the Expectation-step, and update the model parameters Θ. The parameters can be decoupled into two sub-groups. The first sub-group is $\mathbf{B}_j(t), j = 1,\ldots,P$ and $\mathbf{r}$, which are directly related to the observations $\mathcal{Y}$. The second sub-group contains the important $\mathbf{C}$ and ω_0^2, which determine the causal relationship of different nodes, namely, the DAG structure.

For the first sub-group: denote $F_n(\mathbf{B},\Theta')$ as the part of the quadratic loss in Eq.

(11) related to $\mathbf{B}$. Minimizing F_n is equivalent to maximizing Q_n respecting to $\mathbf{B}$:

$$\begin{aligned}\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2, \ldots, \hat{\mathbf{B}}_P \doteq & \underset{\mathbf{B}_1,\ldots,\mathbf{B}_P}{\arg\min} \sum_{n=1}^{N}\sum_{j=1}^{P}\sum_{l=1}^{L_j} \mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta'} \left((\mathbf{Y}_{jl}^{(n)} - \mathbf{B}_j \mathbf{x}_{jl}^{(n)})^T r_{jl}^{-2} (\mathbf{Y}_{jl}^{(n)} - \mathbf{B}_j \mathbf{x}_{jl}^{(n)}) \right) \\ & \propto \frac{1}{N} \sum_{n=1}^{N}\sum_{j=1}^{P}\sum_{l=1}^{L_j} \left(\|\mathbf{Y}_{jl}^{(n)} - \mathbf{B}_j \hat{\mathbf{u}}_{jl;\Theta',\mathbf{Y}^{(n)}}\|_2^2 + \mathrm{tr}(\mathbf{B}_j \hat{\mathbf{\Sigma}}_j \mathbf{B}_j^T) \right) \\ & \doteq F_n(\mathbf{B}, \Theta') \\ & \text{s.t.} \quad \mathbf{B}_j^T \mathbf{B}_j = \mathbf{I} \quad \forall j = 1, \ldots, P. \end{aligned} \tag{17}$$

To solve Eq. (17), we can utilize the polar decomposition. We first calculate $\mathbf{A} = \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L_j} \mathbf{Y}_{jl}^{(n)} \hat{\mathbf{u}}_{jl;\Theta',\mathbf{Y}^{(n)}}^T$, and then perform the polar decomposition on $\mathbf{A}$ to obtain $\mathbf{A} = \mathbf{V}\hat{\mathbf{B}}_j$, where $\mathbf{V}$ is a symmetric matrix, and $\hat{\mathbf{B}}_j$ is the matrix we are interested in.

For estimating $\hat{\mathbf{r}}$, it can be solved in a closed form as:

$$\hat{r}_{jl}^2 = \frac{1}{N}\sum_{n=1}^{N} \left((\mathbf{Y}_{jl}^{(n)} - \hat{\mathbf{B}}_j \hat{\mathbf{u}}_{jl;\Theta',\mathbf{Y}^{(n)}})(\mathbf{Y}_{jl}^{(n)} - \hat{\mathbf{B}}_j \hat{\mathbf{u}}_{jl;\Theta',\mathbf{Y}^{(n)}})^T + \hat{\mathbf{B}}_j \hat{\mathbf{\Sigma}}_j \hat{\mathbf{B}}_j^T) \right), \forall j = 1, \ldots, P, l = 1, \ldots, L_j. \tag{18}$$

For the second sub-group: the key is to infer C, which represents the structure of DAG. Denote G_n as the loss part in $Q_n(\Theta, \Theta')$ related to $\mathbf{C}$. Maximizing Eq. (11) is equivalent to the following:

$$\begin{aligned}\hat{\mathbf{C}}^K, \hat{\mathbf{C}}^L = & \underset{\mathbf{C}^K,\mathbf{C}^L}{\arg\min} \sum_{n=1}^{N}\sum_{j=1}^{P} \mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta'} \left((\mathbf{x}_j^{(n)} - \mathbf{x}^{(n)}\mathbf{C}_j)^T \omega_0^{-2} (\mathbf{x}_j^{(n)} - \mathbf{x}^{(n)}\mathbf{C}_j) \right) + \lambda \|\mathbf{C}\|_{l_1/F} \\ & \propto \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta'} \|\mathbf{x}^{(n)} - \mathbf{x}^{(n)}\mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_{l_1/F} \\ & = \frac{1}{N}\sum_{n=1}^{N} \left(\|\hat{\mathbf{u}}_{\Theta',\mathbf{Y}^{(n)}} - \hat{\mathbf{u}}_{\Theta',\mathbf{Y}^{(n)}}\mathbf{C}\|_2^2 + \mathrm{tr}((\mathbf{I} - \mathbf{C})^T \hat{\mathbf{\Sigma}}_{\Theta'} (\mathbf{I} - \mathbf{C})) \right) + \lambda \|\mathbf{C}\|_{l_1/F} \\ & \doteq G_n(\mathbf{C}, \Theta') + \lambda \|\mathbf{C}\|_{l_1/F} \\ & \text{s.t.} \quad h(\mathbf{W}) = 0, \end{aligned} \tag{19}$$

Algorithm 1 EM algorithm

Input: data $\mathcal{Y}$, tolerances ϵ_0

Initialize $\Theta^{(0)} = \text{vec}(\mathbf{B}^{(0)}, \mathbf{C}^{K(0)}, \mathbf{C}^{L(0)}, \mathbf{R}^{(0)}, \omega^{2(0)})$, $s = 0$.

repeat

$s \leftarrow s + 1$

$\hat{\mathbf{u}}^{(n)}_{\Theta^{(s-1)}, \mathbf{Y}^{(n)}}, \hat{\mathbf{\Sigma}}_{\Theta^{(s-1)}} \leftarrow$ Forward filtering & backward smoothing($\Theta^{(s-1)}$) via Appx. B.3

$\mathbf{B}^{(s)} \leftarrow \arg\min_{\mathbf{B}} F_n(\mathbf{B}, \Theta^{(s-1)})$ via Polar decomposition.

$\mathbf{C}^{K(s)}, \mathbf{C}^{L(s)} \leftarrow \arg\min_{\mathbf{C}^K, \mathbf{C}^L} G_n(\mathbf{C}; \Theta^{(s-1)}) + \lambda \|\mathbf{C}\|_{l_1/F}$ via Algorithm 2

Update $\omega_0^{2(s)}$ by Eq. (21)

$\Theta^{(s)} \leftarrow \text{vec}(\mathbf{B}^{(s)}, \mathbf{C}^{K(s)}, \mathbf{C}^{L(s)}, \mathbf{R}^{(s)}, \omega_0^{2(s)})$

until $D(\Theta^{(s)}, \Theta^{(s-1)}) < \epsilon_0$

where $\mathbf{C}^K$ is a $\sum_j K_j \times \sum_j K_j$ matrix with its (j, j') block as $\mathbf{C}^K_{(j,j')} = \mathbf{C}^K_{j'j}$, $\mathbf{C}^L$ is a $L \times L$ matrix with its (j, j') block as $\mathbf{C}^L_{(j,j')} = \mathbf{C}^L_{j'j}$, $\mathbf{C}_{j'j} = \mathbf{C}^L_{j'j} \otimes \mathbf{C}^K_{j'j}$.

We can convert Eq. (19) into an unconstrained problem using the Lagrangian dual method:

$$\hat{\mathbf{C}}^K, \hat{\mathbf{C}}^L \in \arg\min_{\mathbf{C}^K, \mathbf{C}^L} \max_{b>0} \tilde{G}_n(\mathbf{C}, \Theta') + \lambda \|\mathbf{C}\|_{l_1/F}, \tag{20}$$

where

$$\tilde{G}_n(\mathbf{C}, \Theta') = G_n(\mathbf{C}, \Theta') + bh(\mathbf{W}) + \frac{a}{2} h(\mathbf{W})^2.$$

$b \in \mathbb{R}$ is dual variable and $a \in \mathbb{R}$ is the coefficient for quadratic penalty. We solve the Lagrangian dual problem by the dual ascent method. Due to the non-smoothness of l_1/F norm, we use the proximal gradient method for group lasso penalty. We summarize the algorithm in Algorithm 2.

After obtaining the transition matrix $\hat{\mathbf{C}}$, $\hat{\omega}_0^2$ can be solved in a closed form as

$$\hat{\omega}_0^2 = \frac{1}{NM}\sum_{n=1}^{N}\left(\|\hat{\mathbf{u}}_{\Theta',\mathbf{Y}^{(n)}} - \hat{\mathbf{u}}_{\Theta',\mathbf{Y}^{(n)}}\hat{\mathbf{C}}_j\|_2^2 + \mathrm{tr}((\mathbf{I}-\hat{\mathbf{C}})^T\hat{\mathbf{\Sigma}}_{\Theta'}(\mathbf{I}-\hat{\mathbf{C}}))\right). \tag{21}$$

Combine Eqs. (17), (18), (19) and (21), we can update Θ and replace Θ' by the updated Θ.

We repeat the Expectation-step and Maximization-step iteratively until convergence, i.e., the difference between the estimated parameters

$$D(\Theta,\Theta') = \sqrt{\|\mathbf{C}-\mathbf{C}'\|_F^2 + \|\mathbf{B}-\mathbf{B}'\|_F^2 + \|\mathbf{r}-\mathbf{r}'\|_2^2 + \|\omega_0^2 - \omega_0^{2'}\|_F^2}$$

is smaller than a threshold ϵ_0. We summarize the regularized EM algorithm in Algorithm 1, where $\Theta = \{\mathbf{C},\mathbf{B},\mathbf{r},\omega_0^2\}$ and $\Theta' = \{\mathbf{C}',\mathbf{B}',\mathbf{r}',\omega_0^{2'}\}$.

4 Theoretical Properties

In the following, we prove that when certain model assumptions hold, the estimated parameters can converge to those of the true model. Assuming for the true model, its parameter set is denoted as $\Theta^* = \{\mathbf{C}^*,\mathbf{B}^*,\omega_0^{2*},\mathbf{r}^*\}$. Condition 2 gives the upper and lower bounds of the variances that ensure the data covariance matrix is not degenerate, i.e.,

Condition 2. *All the eigenvalues of* $\mathbf{\Sigma}^* = (\mathbf{I}-\mathbf{C}^*)^{-T}\omega_0^{2*}(\mathbf{I}-\mathbf{C}^*)^{-1}$ *should be greater than a constant* $\eta_{\mathbf{\Sigma}^*} > 0$ *and finite, and* $\forall j \in 1,\dots P$ *and* $l = 1,\dots,L_j$*, we assume* $r_{jl}^2 < \infty$*.*

Condition 3 ensures the identifiability of decomposition on $\mathbf{Y}$.

Condition 3. *The number of latent variables is smaller than the number of sampling points for each function, i.e.,* $K_j < T, \forall j = 1,\dots,P$*.*

By combining these three conditions, we can obtain the good property for all the models in the equivalent class $\mathfrak{D}$ in Theorem 1.

Algorithm 2 Algorithm for Largrangian dual problem

Input: posterior distribution $\hat{\mathbf{u}}^{(n)}_{\Theta^{(s-1)},\mathbf{Y}^{(n)}}, \hat{\mathbf{\Sigma}}_{\Theta^{(s-1)}}$, tolerance h_{tol}, learning rate lr, γ.

Initialize $\mathbf{C}^K, \mathbf{C}^L, a \leftarrow 1, b \leftarrow 0$.

repeat

Update $\mathbf{C}^K, \mathbf{C}^L$ by minimizing $\tilde{G}_n$ by gradient method.

$b \leftarrow b + ah(\mathbf{W})$

$a \leftarrow lr * a$

until $h(\mathbf{W}) < h_{tol}$

for i from 1 to P **do**

for j from 1 to P **do**

if $\|\mathbf{C}_{ij}\|_F > \gamma\lambda$ **then**

$\mathbf{C}^L_{ij} \leftarrow \mathbf{C}^L_{ij} - \gamma\lambda \frac{\mathbf{C}^L_{ij}}{\|\mathbf{C}_{ij}\|_F}$

else

$\mathbf{C}^L_{ij} \leftarrow \mathbf{0}$

end if

end for

end for

Theorem 1 (Equivalence class)**.** *Define the equivalence class of the true parameters* Θ^* *as* $\mathfrak{D}$*. Under Conditions 1 to 3, for any parameters* $\Theta^e = \{\mathbf{C}, \mathbf{B}, \mathbf{r}, \omega_0^2\} \in \mathfrak{D}$*, it can be represented by the following form:*

$$\mathbf{B}_j = \mathbf{B}^*_j \mathbf{Q}_j,$$

$$\mathbf{r}_{jl} = \mathbf{r}^*_{jl},$$

$$\omega_0^2 = \omega_0^{2*},$$

$$\mathbf{C}_{j'jl'l} = \mathbf{Q}_j \mathbf{C}^*_{j'jl'l} \mathbf{Q}^T_{j'},$$

where $\mathbf{Q}_j \in \mathbb{R}^{K_j \times K_j}$ *is an orthogonal matrix satisfying* $\mathbf{Q}_j^T \mathbf{Q}_j = \mathbf{Q}_j \mathbf{Q}_j^T = \mathbf{I}$. *This states that the equivalence class of the true solution is only the set of orthogonal transformations of* Θ^*.

Proof. The proof is in Appx. A.1 □

Intuitively speaking, this indicates though we choose different orthogonal basis functions to map $\mathbf{Y}_{jl}$, the spaces spanned by these orthogonal functional basis spaces are the same. Therefore, we can obtain the true causal structure once we get any equivalent solution $\Theta^e \in \mathfrak{D}$.

Next, we aim to prove that our regularized EM algorithm is capable of discovering the true causal order and parameters when the initial parameters $\Theta^{(0)}$ are close to the true parameters Θ^*. We give the definition of population analogs of F_n and G_n in Definition 1.

Definition 1 (Population analogs)**.** *Define* F *and* G *as the population analogs of* F_n *and* G_n *respectively, i.e.,*

$$F(\mathbf{B}, \Theta') = \int \sum_{j=1}^{P} \sum_{l=1}^{L_j} \mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta'} \|\mathbf{Y}_{jl} - \mathbf{B}_j \mathbf{x}_{jl}\|_2^2 p(\mathbf{Y}; \Theta^*) \mathrm{d}\mathbf{Y},$$

$$G(\mathbf{C}, \Theta') = \int \mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta'} \|\mathbf{x} - \mathbf{x}\mathbf{C}\|_2^2 p(\mathbf{Y}; \Theta^*) \mathrm{d}\mathbf{Y}.$$

Using the Strong Law of Large Numbers, we can observe that as n *approaches infinity, the results* F_n *and* G_n *converge almost surely to* F *and* G *respectively.*

Theorem 2 shows the true parameters Θ^* can maximize the population log-likelihood function and satisfy the self-consistency property (McLachlan and Krishnan, 2007).

Theorem 2 (Self-consistency)**.** *When Conditions 1 and 2 hold, we can obtain* Θ^* *by minimizing* $G(\cdot, \Theta^*)$ *and* $F(\cdot, \Theta^*)$.

Proof. The proof is in Appx. A.2. □

Next, we introduce the theorem related to causal structure. We define the causal order π in Definition 2.

Definition 2. *Since* $\mathbf{W}$ *is the adjacency matrix of a DAG* $\mathcal{G}$*, the nonzero entries of* $\mathbf{W}$ *define the causal order of graph* $\pi \in \mathbb{S}_P$*, which can be represented by a permutation over* $1, 2, ..., P$*.* $\pi(i)$ *represents the position of node* i *in the order. A causal order* π *is consistent with a DAG* $\mathcal{G}$ *if and only if:*

$$\mathbf{W}_{ij} \neq 0 \Rightarrow \pi(i) < \pi(j). \tag{22}$$

With abusive use of notation, we denote $\mathbf{C}(\pi)$ to address this $\mathbf{C}$ is consistent with causal order π. Then define $\mathcal{C}(\pi)$ as the set of $\mathbf{C}(\pi)$ that has the same causal order π, i.e., $\mathbf{C}(\pi) \in \mathcal{C}(\pi)$. Denote $\mathbf{C}^*_\Theta(\pi) = \underset{\mathbf{C}(\pi)\in\mathcal{C}(\pi)}{\arg\min} G(\mathbf{C}(\pi), \Theta)$. Let Π^*_0 be the set of all causal orders consistent with $\mathbf{C}^*$. Since Theorem 2 holds, we have $\mathbf{C}^*_{\Theta^*}(\pi_0) = \mathbf{C}^*, \forall \pi_0 \in \Pi^*_0$.

Condition 4 (Omega-min)**.** *Under Conditions 1 and 2, for all* $\pi \notin \Pi^*_0, \exists \eta_1 > 0$ *that:*

$$G(\mathbf{C}^*, \Theta^*) - G(\mathbf{C}^*_{\Theta^*}(\pi), \Theta^*) < -\eta_1. \tag{23}$$

Condition 4 assumes that if we restrict our model to a wrong causal order $\pi' \notin \Pi^*_0$, $G(\mathbf{C}^*_{\Theta^*}(\pi'), \Theta^*)$ will increase by at least η_1. This is similar to the Omega-min condition in Van de Geer and Bühlmann (2013), and is used to justify the precision of our true model.

Lemma 1. *Under Condition 2 and 3,* $\exists \tilde{r}_1$*, the following inequalities hold for* $\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r}_1)$*:*

(1) $\max\limits_{\Theta\in\mathbb{B}_2(\Theta^*,\tilde{r}_1)} \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\|\mathbf{x}\|_2^8) < \infty$*;*

(2) $\max\limits_{\Theta\in\mathbb{B}_2(\Theta^*,\tilde{r}_1)} \max\limits_{j,l} \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\|\mathbf{Y}_{jl}\hat{\mathbf{u}}^T_{jl,\Theta,\mathbf{Y}}\|_F^4) < \infty$*;*

(3) $\min\limits_{\Theta\in\mathbb{B}_2(\Theta^*,\tilde{r}_1)} \text{minEig}(\text{Cov}(\hat{\mathbf{u}}_{\Theta^*,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_{\Theta^*}) > 0$*;*

(4) $\min\limits_{\Theta\in\mathbb{B}_2(\Theta^*,\tilde{r}_1)} \min\limits_{j,l} \sigma_{\min}(\mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\mathbf{Y}_{jl}\hat{\mathbf{u}}^T_{jl,\Theta,\mathbf{Y}})) > 0$*;*

where $\text{minEig}(\cdot)$ *is the minimum eigenvalue of the matrix.* $\sigma_{\min}(\mathbf{A})$ *is* k*-th maximum singular value of the matrix for* $\mathbf{A} \in \mathbb{R}^{T\times k}$*, where* $\sigma_{\min}(\cdot) > 0$ *shows that the matrix is column full rank.* $\mathbb{B}_2(\Theta^*, r) := \{\Theta | D(\Theta, \Theta^*) \leq r\}$.

Proof. The proof is in Appx. A.3. □

Lemma 1 shows that the posterior distribution $p(\mathbf{x}|\mathbf{Y};\Theta)$ is not degraded and has bounded variance when $\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r}_1)$. We denote

$$\sup_{\Theta\in\mathbb{B}_2(\Theta^*,\tilde{r}_1)} \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\|\mathbf{x}\|_2^8) = \mathtt{x}_{\sup}^8,$$

$$\sup_{\Theta\in\mathbb{B}_2(\Theta^*,\tilde{r}_1)} \sup_{j,l} \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\|\mathbf{Y}_{jl}\hat{\mathbf{u}}_{jl,\Theta,\mathbf{Y}}^T\|_F^4) = \mathtt{y}_{\sup}^4,$$

$$\inf_{\Theta\in\mathbb{B}_2(\Theta^*,\tilde{r}_1)} \text{minEig}(\text{Cov}(\hat{\mathbf{u}}_{\Theta,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_\Theta) = \mathtt{s}_{\inf},$$

$$\inf_{\Theta\in\mathbb{B}_2(\Theta,\tilde{r}_1)} \min_{j,l} \sigma_{\min}(\mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x};\Theta}(\mathbf{Y}_{jl}\hat{\mathbf{u}}_{jl,\Theta,\mathbf{Y}}^T)) = \mathtt{b}_{\inf},$$

where $\mathtt{x}_{\sup}^8, \mathtt{y}_{\sup}^4, \mathtt{s}_{\sup}, \mathtt{b}_{\inf} > 0$ are universal constants depended on $\tilde{r}_1$.

Lemma 2. *Under Condition 4,* $\exists\tilde{r}_2, \forall\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r}_2)$*, denote* $\Pi_\Theta^* = \{\pi | \pi = \arg\min_{\pi'} G(\mathbf{C}_\Theta^*(\pi'), \Theta)\}$ *and* $\mathbf{C}_\Theta^* = \underset{\mathbf{C}_\Theta^*(\pi)}{\arg\min} G(\mathbf{C}_\Theta^*(\pi);\Theta)$*. We have:*

(1) $\Pi_\Theta^* = \Pi_0^*$*,*

(2) For all $\pi \notin \Pi_\Theta^*, \exists 0 < \eta_2 < \eta_1$ *that:*

$$G(\mathbf{C}_\Theta^*, \Theta) - G(\mathbf{C}_\Theta^*(\pi), \Theta) < -\eta_2. \tag{24}$$

Proof. The proof is in Appx. A.4. □

Lemma 2 extends Condition 4 from Θ^* to all $\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r}_2)$. It states that when Θ is close to Θ^*, we can still identify the true causal order by minimizing $G(\mathbf{C}, \Theta)$. Taking $\tilde{r} = \min(\tilde{r}_1, \tilde{r}_2)$, Lemma 3 and 4 provide the lower bound for the error when estimating $\hat{\mathbf{C}}$ and $\hat{\mathbf{B}}$, in a single iteration of the regularized EM iteration.

Lemma 3. *Under Conditions 1, 2, 4 and suppose that we solve the optimization of Eq. (19) with specified regularization parameters λ and $\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r})$. Since $\mathbb{B}_2(\Theta^*, \tilde{r})$ is a contact set, we denote $\mathtt{c}_{\sup} \doteq \sup_{\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r})} \sup_{\pi} \|\mathbf{C}^*_{\Theta}(\pi)\|_{l_1/F}$ and $\mathtt{d}^4_{\sup} = \sup_{\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r})} \sup_{\pi} \|\mathbf{I} - \mathbf{C}^*_{\Theta}(\pi)\|^4_F$. If the following conditions are satisfied for $\varrho_1, \varrho_2, \varrho_3 \in (0,1), \delta_1 \in (0, 1/2)$:*

$$\eta_2 > 2\sqrt{\frac{\mathtt{d}^4_{\sup}\mathtt{x}^4_{\sup}}{\varrho_1 N}} + \lambda(2\delta_1 + 1)\mathtt{c}_{\sup},$$
$$\frac{\mathtt{d}^2_{\sup}\mathtt{x}^4_{\sup}}{\lambda^2 N \delta_1^2} < 1,$$
$$1 - 2\varrho_1 - P!M\varrho_2 - \varrho_3 > 0,$$
$$\mathtt{s}_{\inf} > \sqrt{\frac{\mathtt{x}^4_{\sup}}{N} + \sqrt{\frac{\mathtt{x}^8_{\sup}}{\varrho_3 N}}}.$$

Denote $\hat{\mathbf{C}}$ and $\hat{\pi}$ as the matrix and corresponding causal order by solving Eqs. (19) with $\Theta' = \Theta$. Then the following statements hold true:

(1) With probability at least $1 - 2\varrho_1 - P!M\varrho_2$, $\hat{\pi} \in \Pi^*_0$;

(2) With probability at least $1 - 2\varrho_1 - P!M\varrho_2 - \varrho_3$,

$$\|\hat{\mathbf{C}} - \mathbf{C}^*_{\Theta}\|^2_F \leq \frac{2\sqrt{\frac{\mathtt{d}^4_{\sup}\mathtt{x}^4_{\sup}}{\varrho_1 N}} + \lambda(2\delta_1 + 1)\mathtt{c}_{\sup}}{\mathtt{s}_{\inf} - \sqrt{\frac{\mathtt{x}^4_{\sup}}{N} + \sqrt{\frac{\mathtt{x}^8_{\sup}}{\varrho_3 N}}}}.$$

Proof. The proof is in Appx. A.5. □

Lemma 4. *Under Condition 2 and 3, denote $\mathbf{B}^*_{\Theta}$ as the matrix that minimizes $F(\cdot, \Theta)$ with $\mathbf{B}^{T*}_{\Theta j}\mathbf{B}^*_{\Theta j} = \mathbf{I}, \forall j$ and $\hat{\mathbf{B}}$ as the optimal solution to $F_n(\cdot, \Theta)$ with $\hat{\mathbf{B}}^T_j \hat{\mathbf{B}}_j = \mathbf{I}, \forall j$. Then if for $\varrho_4, \varrho_5 \in (0,1)$:*

$$1 - P\varrho_4 - P\varrho_5 > 0, \quad \text{and} \quad \mathtt{b}_{\inf} - \sqrt{\frac{\mathtt{y}^2_{\sup}}{N} + \sqrt{\frac{\mathtt{y}^4_{\sup}}{\varrho_5 N}}} > 0,$$

with probability at least $1 - P\varrho_4 - P\varrho_5$, we have:

$$\|\hat{\mathbf{B}} - \mathbf{B}^*_{\Theta}\|^2_F \leq \frac{P\left(\frac{\mathtt{y}^2_{\sup}}{N} + \sqrt{\frac{\mathtt{y}^4_{\sup}}{\varrho_4 N}}\right)}{\left(\mathtt{b}_{\inf} - \sqrt{\frac{\mathtt{y}^2_{\sup}}{N} + \sqrt{\frac{\mathtt{y}^4_{\sup}}{\varrho_5 N}}}\right)^2}.$$

Proof. The proof is in Appx. A.6. □

Next, we aim to derive an upper bound for the total error bound of our regularized EM algorithm, i.e., $D(\Theta^{(S)}, \Theta^*)$ for total S iterations in the regularized EM algorithm. Under certain conditions (seeing Conditions 5 and 6 in Appendix. B.2), Theorem 3 and Corollary 1 establish the convergence properties and error analysis of our regularized EM algorithm. These results hold when the initial solution $\Theta^{(0)}$ is in proximity to the true solution, encompassing scenarios of both finite N and as N approaches infinity. Furthermore, this property still holds when replacing Θ^* with any equivalent solution $\Theta^e \in \mathfrak{D}$. As Theorem 1 states, any $\Theta^e \in \mathfrak{D}$ has the same causal structure as Θ^*. Therefore, we show that our regularized EM algorithm can effectively learn the correct causal structure locally.

Theorem 3. *Assume Conditions 1 to 6 and the conditions in Lemmas 3 and 4 are satisfied, and the EM estimator* $M(\Theta) \doteq \arg\max_{\Theta'} Q(\Theta'; \Theta) - \lambda \mathcal{R}(\mathbf{C})$ *is contractive with parameters* $\kappa \in (0, 1)$ *in the ball* $\mathbb{B}_2(\Theta^*, \tilde{r})$. *Denote* S *as the total iterations of the regularized EM algorithm, we have*

$$D(\Theta^{(S)}, \Theta^*) \leq \kappa^S D(\Theta^{(0)}, \Theta^*) + \frac{1}{1-\kappa} \epsilon(\delta/S, N/S, \tilde{r}),$$

where

$$\delta/S = 2\varrho_1 + MP!\varrho_2 + \varrho_3 + P\varrho_4 + P\varrho_5 + M\varrho_6,$$

$$\epsilon(\delta/S, N/S, \tilde{r}) = \left(\frac{2\sqrt{\frac{\mathtt{d}_{\sup}^4 \mathtt{x}_{\sup}^4 S}{\varrho_1 N}} + \lambda(2\delta_1 + 1)\mathtt{c}_{\sup}}{\mathtt{s}_{\inf} - \sqrt{\frac{\mathtt{x}_{\sup}^4 S}{N}} + \sqrt{\frac{\mathtt{x}_{\sup}^8 S}{\varrho_3 N}}} + \frac{P\left(\frac{\mathtt{y}_{\sup}^2 S}{N} + \sqrt{\frac{\mathtt{y}_{\sup}^4 S}{\varrho_4 N}} \right)}{\left(\mathtt{b}_{\inf} - \sqrt{\frac{\mathtt{y}_{\sup}^2 S}{N}} + \sqrt{\frac{\mathtt{y}_{\sup}^4 S}{\varrho_5 N}} \right)^2} \right)^{1/2} + O((N/S)^{-1/2}).$$

Proof. According to Lemma 3 and 4 above, together with Lemma 11 and 12 which gives the error bound of $\mathbf{r}$ and ω_0^2 in each EM iteration as $O((N/S)^{-1/2})$ with probability $1 - M\varrho_6$, we can prove Theorem 3 following the procedures in Theorem 5 in Balakrishnan et al. (2017). □

Corollary 1 (Asymptotic property)**.** *Based on Theorem 3, we have the following two corollaries:*

(1) As $N \to \infty$, by setting $\lambda \sim N^{-1/2+\nu}$ with $\nu \in (0, 1/2)$, the conditions in Lemma 3 hold. Then $\epsilon(\delta/S, N/S, \tilde{r}) = \mathcal{O}((N/S)^{(-1+2\nu)/4})$, and the total estimation error after S EM iterations can be described as $D(\Theta^{(S)}, \Theta^*) \leq \kappa^S D(\Theta^{(0)}, \Theta^*) + \mathcal{O}((N/S))^{(-1+2\nu)/4})$.

(2) Under $S \to \infty$, $N \to \infty$ and $N/S \to \infty$, we have $\Theta^{(S)} \to \Theta^*$ with probability 1.

5 Numerical study

To evaluate the performance of our methodology and selection of λ, we apply our MultiFun-DAG to solve a synthetic graphical model. We show the performance of our algorithm on tasks of different combinations (N, P, L_0, K_0), where $\forall j = 1, \ldots, P$, we have $L_j = L_0$ and $K_j = K_0$.

In each experiment, the graphs are generated by Erdös-Rényi random graph model, where the functional data of the different nodes have the same Fourier basis $\nu_1(t)$, $\nu_2(t)$, ..., $\nu_K(t)$:

$$\nu_k(t) = \begin{cases} 1, & k = 1, \\ \cos(2\pi u t), & k = 2u, \\ \sin(2\pi u t), & k = 2u+1, \end{cases} \qquad \forall u \in \mathbb{Z}, u \geq 1.$$

By combining Eq. (1), Eq. (4) and Eq. (8), we can write the representation of each functional data. The generated transition matrix is $\mathbf{C}_{j'j} = c_{j'j} \mathbf{1}_{L_{j'} \times L_j} \otimes \mathbf{I}_K$, where $c_{j'j}$ is independently and identically generated from a uniform distribution $\mathcal{U}(-2, 0.5) \cup (0.5, 2)$. The variance of noise is set by $\omega_0^2 = 1$ and $r_{jl}^2 = 0.01, \forall j = 1, \ldots, P$.

For model comparison, we select two methods from the literature and another two variants of our MultiFun-DAG. The baselines compared in this paper are introduced below.

Since they cannot be directly used for DAG with nodes as multivariate functions, we modify these methods by concatenating multivariate functions as long univariate functions for analysis.

- **FDGM_S**: The functional directed graph model proposed by Sun et al. (2017). To deal with multivariate functional data for each node, we concatenate L_j functional data of each node as long functional data with $L_j * T$ observation points.

- **FDGM_G**: The functional directed graph model proposed by Gómez et al. (2020). To deal with multivariate functional data for each node, we concatenate L_j functional data as long univariate functional data with $L_j * T$ observation points.

- **MFGM**: This baseline provides a two-stage method to model the multivariate functional DAG. It first implements FPCA for each node separately to obtain their PC scores. Then it treats these scores as $\mathbf{X}$ and uses the same structural learning method as MultiFun-DAG to estimate the causal structure, i.e.,

$$\min_{\mathbf{C}^K, \mathbf{C}^L} \frac{1}{N} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \lambda \|\mathbf{C}\|_{l_1/F},$$
$$\text{s.t.} \quad \text{tr}(\exp(\mathbf{W} \circ \mathbf{W})) - P = 0,$$

  where $\mathbf{W}$ and λ have the same meaning as our method.

- **NoTears**: It first implements FPCA for the functional data of each node, where all the nodes share a common set of K bases. After FPCA, the causal relationships between each PC score of each original node are learned by NoTears (Zheng et al., 2018). Then the causal relationships between all the PC scores from two nodes are merged as the final causal relationship between these two nodes.

In this experiment, we aim to test the effectiveness of different methods to recover the true DAG structures. For brevity, we give the F1 score of the arcs to represent model

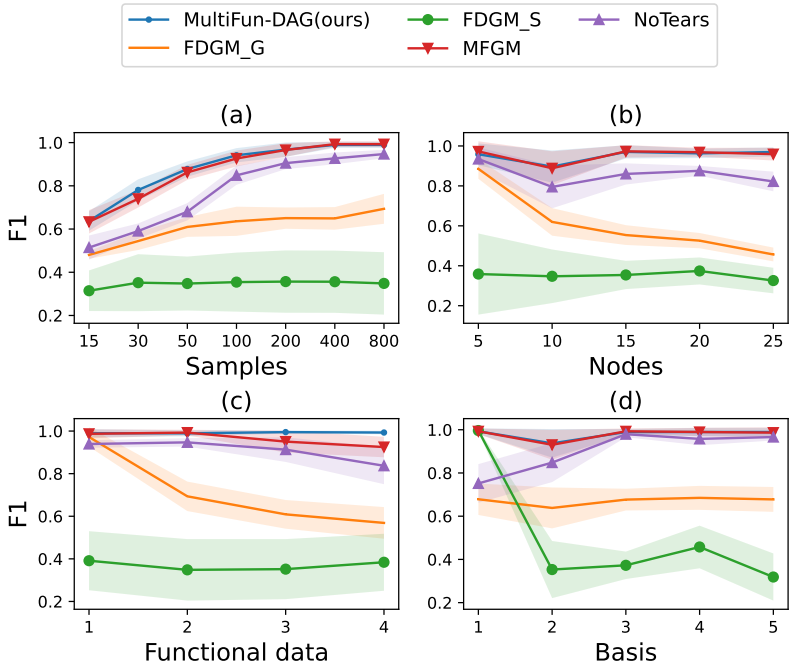

Figure 3: F1 score of the edges with 95% confidence intervals: F1 score with different numbers of (a) samples N; (b) nodes P; (c) functions L_0; (d) bases K_0.

performance, i.e.,

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

In Fig. 3, we see that our MultiFun-DAG has the best performance among all the baselines. The performance increases as the number of samples N increases. MFGM has a similar performance to MultiFun-DAG but performs worse when the number of function data L_0 increases. This justifies the importance of our joint estimation of $\mathbf{X}$ and $\mathbf{C}$.

Meanwhile, by comparing MFGM with NoTears, we verify the benefit of learning the DAG with vector-value nodes over the DAGs with scalar-value nodes. The difference in performance between MFGM and NoTears increases as the number of nodes increases or the number of functions increases. This is due to model complexity, i.e., the search space of causal order in NoTears is much larger than that in MFGM.

Table 1 compares the result of our estimated parameters and the true parameters. In

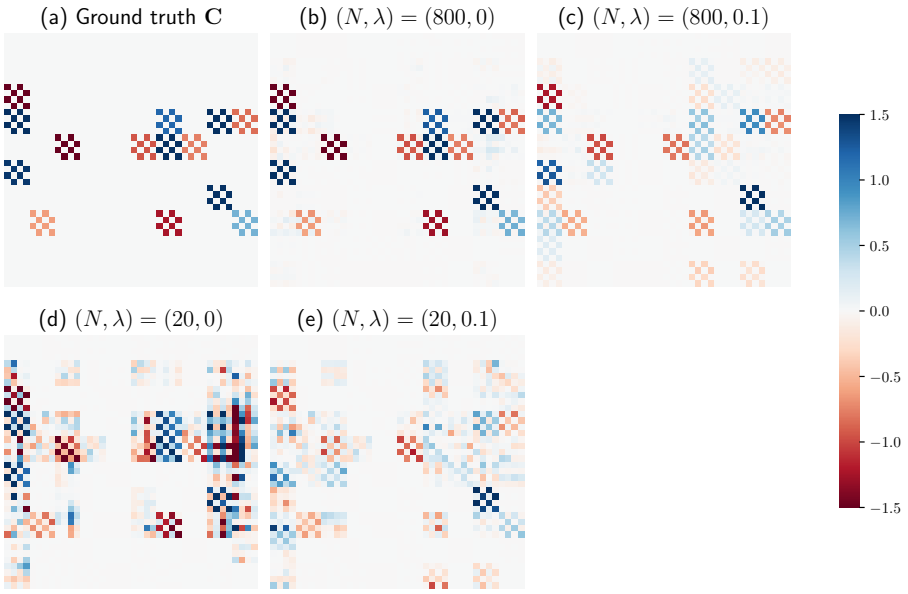

Figure 4: Heatmap of $\mathbf{C}^*$ and the estimated $\tilde{\mathbf{C}}$ by MultiFun-DAG. Titles of the subplots represent the results under different experiment settings of (N, λ).

this case, we rotate the matrix $\hat{\mathbf{B}}$ to the true matrix $\mathbf{B}^*$. The rotation equation is given by Theorem 1: $\mathbf{B}_j^* = \hat{\mathbf{B}}_j \mathbf{Q}_j$ and $\tilde{\mathbf{C}}_{jk} = \mathbf{C}_{jk}^L \otimes (\mathbf{Q}_j^T \mathbf{C}_{jk}^K \mathbf{Q}_k)$, where $\mathbf{Q}_j$ is an orthogonal matrix. The rotation process maintains the structure of DAG. Then we compare $\tilde{\mathbf{C}}$ and $\mathbf{C}^*$, by $\|\tilde{\mathbf{C}} - \mathbf{C}^*\|_F^2$. Furthermore, MSE_{est} and MSE_{true} measures the l_2 loss of $\mathcal{Y}$ for the estimated model and the true model, which can be computed by:

$$\text{MSE}_{\text{est}} = \frac{1}{NLT} \sum_{n=1}^{N} \sum_{j=1}^{P} \sum_{l=1}^{L_j} \mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y},\hat{\Theta}} \|\mathbf{Y}_{jl}^{(n)} - \hat{\mathbf{B}} \hat{\mathbf{x}}_{jl}^{(n)}\|_2^2,$$

$$\text{MSE}_{\text{true}} = \frac{1}{NLT} \sum_{n=1}^{N} \sum_{j=1}^{P} \sum_{l=1}^{L_j} \|\mathbf{Y}_{jl}^{(n)} - \mathbf{B}^* \mathbf{x}_{jl}^{(n)}\|_2^2.$$

When $\text{MSE}_{\text{est}} < \text{MSE}_{\text{true}}$, overfitting occurs. When $\text{MSE}_{\text{est}} > \text{MSE}_{\text{true}}$, underfitting occurs. A smaller $|\text{MSE}_{\text{est}} - \text{MSE}_{\text{true}}|$, which is denoted by $|\Delta|$, indicates a smaller difference between the estimated and true parameters. Besides, a smaller N needs a larger λ to prevent overfitting, and on the contrary, a larger N needs a smaller λ to prevent underfitting.

Table 1: Estimated parameters v.s. True parameters.

N	λ	$\|\tilde{\mathbf{C}} - \mathbf{C}^*\|_F^2$	MSE_{est}	MSE_{true}	$\|\Delta\|$
800	0	1.21	1.99	2.013	0.02
800	0.1	47.20	2.29	2.013	0.28
20	0	238.40	0.99	2.006	1.02
20	0.1	107.74	1.46	2.006	0.55

This might be because large λ increases the bias and robustness of our algorithm. Fig. 4 visualizes the estimated $\tilde{\mathbf{C}}$ (the structure) under different experiment scenarios and its ground truth. With $(N, \lambda) = (800, 0)$, we could faithfully recover the structure.

6 Case study

In this section, we illustrate how our method can be applied to real-world urban traffic data for root cause analysis of traffic congestion. We focus on three types of traffic variables (nodes). (1) The real-time **traffic setting variables**, such as the real-time Origin-Destination (OD) demand, turning probability, the cycle time of the traffic light, etc., denoted as $\mathbf{S}(t) = [\mathbf{S}_1(t), \mathbf{S}_2(t), ..., \mathbf{S}_{P_s}(t)]$. (2) The real-time **traffic condition variables**, such as the occupancy of each lane, the average speed of each lane, the average waiting time of each lane, the number of vehicles in each lane, the number of halting vehicles in each lane, etc., denoted as $\mathbf{Y}(t) = [\mathbf{Y}_1(t), \mathbf{Y}_2(t), ..., \mathbf{Y}_{P_y}(t)]$. (3) The real-time traffic **congestion root cause variables**, such as long/short cycle time of traffic lights, phase imbalance, irrational guide lane, irrational phase sequence, imbalance of entrance, etc., denoted as $\mathbf{R}(t) = [\mathbf{R}_1(t), \ldots, \mathbf{R}_{P_r}(t)]$. Table 2 summarizes the abbreviations and descriptions of each node.

We use the Simulation of Urban MObility (SUMO) (Krajzewicz et al., 2002) to syn-

thesize the real-time traffic data. We collect data from $\mathbf{S}(t)$ and $\mathbf{Y}(t)$ every five minutes and simulate for 60 minutes. Therefore, each functional data has $T = 12$ observation points. For each node of $\mathbf{Y}_j(t), t = 1, \ldots, T$, it has four functions, defined as $\mathbf{Y}_j \in \mathbb{R}^{4\times T}, j = 1, \ldots, P_y$. For each node of $\mathbf{S}_j(t)$ and $\mathbf{R}_j(t)$, it is a univariate function, defined as $\mathbf{S}_j \in \mathbb{R}^T, j = 1, \ldots, P_s$ and $\mathbf{R}_j \in \mathbb{R}^T, j = 1, \ldots, P_r$. We set $\mathbf{R}_j(t) \in \{0, 1\}$. Here $\mathbf{R}_j(t) = 1$ indicates that the j-th type of congestion appears at time t, which is decided by rule-based algorithms in transportation. Its data is also collected every five minutes, with the same sampling grids as the other two types of traffic variables.

Table 2: Abbreviation and the description of traffic data

Node	Name	Description
$\mathbf{S}_1 \in \mathbb{R}^T$	OD-A	OD demand of all direction
$\mathbf{S}_2 \in \mathbb{R}^T$	OD-S	OD demand of certain direction
$\mathbf{S}_3 \in \mathbb{R}^T$	T-A	Turning probability of all direction
$\mathbf{S}_4 \in \mathbb{R}^T$	T-S	Turning probability of certain direction
$\mathbf{S}_5 \in \mathbb{R}^T$	CT	Cycle time of traffic light
$\mathbf{Y}_1 \in \mathbb{R}^{4\times T}$	OC	Occupancy of each of 4 lanes
$\mathbf{Y}_2 \in \mathbb{R}^{4\times T}$	MS	Mean speed of each of 4 lanes
$\mathbf{Y}_3 \in \mathbb{R}^{4\times T}$	MW	Mean waiting time of each of 4 lanes
$\mathbf{Y}_4 \in \mathbb{R}^{4\times T}$	NV	# of vehicles in each of 4 lanes
$\mathbf{Y}_5 \in \mathbb{R}^{4\times T}$	NH	# of halting vehicles in each of 4 lanes
$\mathbf{R}_1 \in \mathbb{R}^T$	Cycle-L	Long cycle time of traffic light
$\mathbf{R}_2 \in \mathbb{R}^T$	Cycle-S	Short cycle time of traffic light
$\mathbf{R}_3 \in \mathbb{R}^T$	Phase-imb	Phase imbalance
$\mathbf{R}_4 \in \mathbb{R}^T$	lanes-irr	Irrational guide lane
$\mathbf{R}_5 \in \mathbb{R}^T$	Entrance-imb	Imbalance of entrance
$\mathbf{R}_6 \in \mathbb{R}^T$	Cycle-irr	Irrational phase sequence

In the experiment, we set 11 levels on $\mathbf{S}_1$, 3 levels on $\mathbf{S}_2$, 3 levels on $\mathbf{S}_3$, 4 levels on $\mathbf{S}_4$ and 4 levels on $\mathbf{S}_5$. Therefore, we have $11 \times 3 \times 3 \times 4 \times 4 = 1584$ treatment combinations. We run a single experiment on each treatment. In each experiment of $\mathbf{S}$, we collect the traffic situation variables $\mathbf{Y}$ and the congestion indicator variables $\mathbf{R}$, and treat them as one sample $[\mathbf{S}^{(n)}, \mathbf{Y}^{(n)}, \mathbf{R}^{(n)}]$ for $n = 1, 2, ..., 1584$.

Then we use MultiFun-DAG to learn the causal relationships between traffic setting variables and traffic congestion root cause variables. Based on domain knowledge, traffic setting variables have effects on the root cause variables, and different types of root cause variables will affect traffic condition variables. Therefore we assume the one-way connection from $\mathbf{S}$ to $\mathbf{R}$ and from $\mathbf{R}$ to $\mathbf{Y}$. Moreover, we assume that there are no interior edges between nodes in $\mathbf{S}$ and nodes in $\mathbf{Y}$. However, we assume that some types of congestion will lead to other types of congestion, i.e., there can be interior edges between nodes in $\mathbf{R}$.

The causal relationships between the variables in MultiFun-DAG are illustrated in Fig. 5, and the probability interpretations are provided. The explainable insights about traffic congestion can be derived. For example, the edges Lanes-irr $\rightarrow$ Phase-imb and Cycle-S indicate that the irrationality of the guide lane could lead to the imbalanced traffic flow in different traffic signal phases, with some directions having long traffic queues and relatively short phase cycle. Thus, the guide lane should be better planned and the cycle time should be extended. In reality, the conditional probability $P(\mathbf{R}_i|\mathbf{S}, \mathbf{Y})$ could also be used to predict the root cause probability in reality.

7 Conclusion

This paper presents a new framework for DAG with nodes as heterogeneous multivariate functional data. It simultaneously conducts functional decomposition for each node and

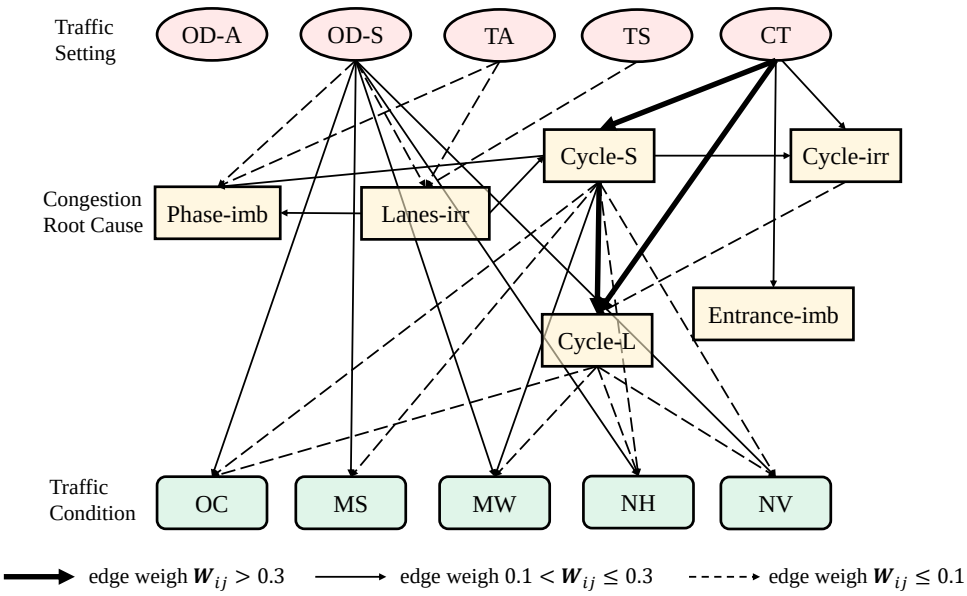

Figure 5: The causal structure of traffic data.

uses the decomposition coefficients to represent the linear causal relationships between different nodes. By conducting a tailored regularized EM algorithm, the DAG structure together with other model parameters can be estimated based on a score-based structural learning algorithm with continuous acyclic constraint. The effectiveness of our algorithm is demonstrated by both theoretical proofs and numerical studies. Some future works include extending the current MultiFun-DAG model to graphs with multi-mode data with both functional nodes and vector nodes. It is also interesting to conduct root causal analysis based on MultiFun-DAG for anomaly detection in multivariate functional data.

References

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., and Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software*, 26(12):1376–

1388.

Aragam, B., Amini, A. A., and Zhou, Q. (2015). Learning directed acyclic graphs with penalized neighbourhood regression. *arXiv preprint arXiv:1511.08963*.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. (2021). Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR.

Chen, Y., Goldsmith, J., and Ogden, R. T. (2018). Functional data analysis of dynamic pet data. *Journal of the American Statistical Association*.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Fraiman, R., Justel, A., Liu, R., and Llop, P. (2014). Detecting trends in time series of functional data: A study of antarctic climate change. *Canadian Journal of Statistics*, 42(4):597–609.

Gómez, A. M. E., Paynabar, K., and Pacella, M. (2020). Functional directed graphical models and applications in root-cause analysis and diagnosis. *Journal of Quality Technology*, 53(4):421–437.

Harris, N. and Drton, M. (2013). Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11).

Heckerman, D. (2008). A tutorial on learning with bayesian networks. *Innovations in Bayesian networks*, pages 33–82.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169.

Krajzewicz, D., Hertkorn, G., Rössel, C., and Wagner, P. (2002). Sumo (simulation of urban mobility)-an open-source traffic simulation. In *Proceedings of the 4th middle East Symposium on Simulation and Modelling (MESM20002)*, pages 183–187.

Lan, T., Li, Z., Li, Z., Bai, L., Li, M., Tsung, F., Ketter, W., Zhao, R., and Zhang, C. (2023). Mm-dag: Multi-task dag learning for multi-modal data - with application for traffic congestion analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1188–1199, New York, NY, USA. Association for Computing Machinery.

Li, B. and Solea, E. (2018). A nonparametric graphical model for functional data with application to brain networks based on fmri. *Journal of the American Statistical Association*, 113(524):1637–1655.

Li, R.-C. (1993). A perturbation bound for the generalized polar decomposition. *BIT Numerical Mathematics*, 33:304–308.

Li, R.-C. (1994). *Relative perturbation theory:(I) eigenvalue variations*. Computer Science Division (EECS), University of California.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59.

Nandy, P., Hauser, A., and Maathuis, M. H. (2018). High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183.

Ng, I., Ghassami, A., and Zhang, K. (2020). On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954.

Qiao, X., Guo, S., and James, G. M. (2019). Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222.

Qiao, X., Qian, C., James, G. M., and Guo, S. (2020). Doubly functional graphical models in high dimensions. *Biometrika*, 107(2):415–431.

Ruz, G. A., Henríquez, P. A., and Mascareño, A. (2020). Sentiment analysis of twitter data during critical events through bayesian networks classifiers. *Future Generation Computer Systems*, 106:92–104.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Sun, H., Huang, S., and Jin, R. (2017). Functional graphical models for manufacturing process modeling. *IEEE Transactions on Automation Science and Engineering*, 14(4):1612–1621.

Van de Geer, S. and Bühlmann, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567.

Velikova, M., van Scheltinga, J. T., Lucas, P. J., and Spaanderman, M. (2014). Exploiting causal functional relationships in bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning*, 55(1):59–73.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015). High dimensional em algorithm: Statistical optimization and asymptotic normality. *Advances in neural information processing systems*, 28.

Wu, H., Zhang, C., and Li, Y.-F. (2022). Monitoring heterogeneous multivariate profiles based on heterogeneous graphical model. *Technometrics*, 64(2):210–223.

Yi, X. and Caramanis, C. (2015). Regularized em algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*, 28.

Zapata, J., Oh, S.-Y., and Petersen, A. (2022). Partial separability and functional graphical models for multivariate gaussian processes. *Biometrika*, 109(3):665–681.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. (2020). Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR.

Appendix

A Proof of theoretical property

A.1 Proof of Theorem 1

Proof. Denote $\Theta_1 := \{\mathbf{C}^{(1)}, \mathbf{B}^{(1)}, \mathbf{r}^{(1)}, \omega_0^{2(1)}\}$ and $\Theta_2 := \{\mathbf{C}^{(2)}, \mathbf{B}^{(2)}, \mathbf{r}^{(2)}, \omega_0^{2(2)}\}$ are two solution in the equivalence class $\mathfrak{D}$. Denote $\mathbf{\Sigma}^{(1)} = (\mathbf{I} - \mathbf{C}^{(1)})^{-T}\omega_0^{2(1)}(\mathbf{I} - \mathbf{C}^{(1)})^{-1}$ and $\mathbf{\Sigma}^{(2)} = (\mathbf{I} - \mathbf{C}^{(2)})^{-T}\omega_0^{2(2)}(\mathbf{I} - \mathbf{C}^{(2)})^{-1}$ are the covariance matrices of $\mathbf{x}$ determined by Θ_1 and Θ_2. Then the following equations hold true:

$$\mathbf{B}_j^{(1)}\mathbf{\Sigma}_{jl,jl}^{(1)}\mathbf{B}_j^{(1)T} + r_{jl}^{2(1)}\mathbf{I}_T = \mathbf{B}_j^{(2)}\mathbf{\Sigma}_{jl,jl}^{(2)}\mathbf{B}_j^{(2)T} + r_{jl}^{2(2)}\mathbf{I}_T \qquad \forall j, l, \tag{25}$$

$$\mathbf{B}_j^{(1)}\mathbf{\Sigma}_{jl,jl}^{(1)}\mathbf{B}_{j'}^{(1)T} = \mathbf{B}_j^{(2)}\mathbf{\Sigma}_{jl,j'l'}^{(2)}\mathbf{B}_{j'}^{(2)T} \qquad \forall (j,l) \neq (j',l'). \tag{26}$$

For the Eq. (25), we have:

$$\mathbf{B}_j^{(1)}\mathbf{\Sigma}_{jl,jl}^{(1)}\mathbf{B}_j^{(1)T} - \mathbf{B}_j^{(2)}\mathbf{\Sigma}_{jl,jl}^{(2)}\mathbf{B}_j^{(2)T} = (r_{jl}^{2(2)} - r_{jl}^{2(1)})\mathbf{I}_T, \tag{27}$$

If $r_{jl}^{2(2)} - r_{jl}^{2(1)} \neq 0$ in Eq. (27), the rank of the right-hand side is T, while the rank of the left-hand side is less than or equal to $K_j < T$, so the equation does not hold. Therefore, we have $r_{jl}^{2(2)} - r_{jl}^{2(1)} = 0$, and $\mathbf{B}_j^{(1)}\mathbf{\Sigma}_{jl,jl}^{(1)}\mathbf{B}_{j'}^{(1)T} = \mathbf{B}_j^{(2)}\mathbf{\Sigma}_{jl,j'l'}^{(2)}\mathbf{B}_{j'}^{(2)T}, \forall j, j', l, l'$. This implies that $\mathbf{B}_j^{(1)} = \mathbf{B}_j^{(2)}\mathbf{Q}_j$ with orthogonal matrix $\mathbf{Q}_j$. From Eq. (26), we obtain $\mathbf{\Sigma}_{jl,j'l'}^{(1)} = \mathbf{Q}_j\mathbf{\Sigma}_{jl,j'l'}^{(2)}\mathbf{Q}_{j'}^T$.

The optimality and uniqueness of the solution are proved in Lemma 5.1 in Aragam et al. (2015) under the assumption of equal variances (Condition 1). It is shown that for any given $\mathbf{\Sigma}^{(1)}$, there exists a unique solution of $\mathbf{C}^{(1)}$. We can show that for any $\mathbf{\Sigma}^{(2)}$ satisfying $\mathbf{\Sigma}_{jl,j'l'}^{(1)} = \mathbf{Q}_j\mathbf{\Sigma}_{jl,j'l'}^{(2)}\mathbf{Q}_{j'}^T$, $\mathbf{C}^{(2)}$ satisfying $\mathbf{Q}_j\mathbf{C}_{j'jl'l}^{(2)}\mathbf{Q}_{j'}^T = \mathbf{C}_{j'jl'l}^{(1)}$ is also the unique solution for $\mathbf{\Sigma}^{(2)}$.

□

A.2 Proof of Theorem 2

Proof. It is equivalent to prove that the optimal points to $F(\cdot, \Theta^*)$ and $G(\cdot, \Theta^*)$ are unique since $\hat{\mathbf{r}}$ and $\hat{\omega}_0^2$ are determined on $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$. The uniqueness of $F(\cdot, \Theta^*)$ is guaranteed by the uniqueness of polar decomposition. As for $G(\cdot, \Theta^*)$, the uniqueness is proved by Lemma 5.1 in Aragam et al. (2015). □

A.3 Proof of Lemma 1

Proof. Proposition. 1 shows that the mean of posterior distribution $\hat{\mathbf{u}}_{\Theta,\mathbf{Y}}$ can be represented by $\hat{\mathbf{u}}_{\Theta,\mathbf{Y}} = \mathbf{A}_\Theta \text{vec}(\mathbf{Y})$ and the covariance is represented by $\hat{\mathbf{\Sigma}}_\Theta$. It is easy to show that $\mathbf{A}_\Theta$ and $\hat{\mathbf{\Sigma}}_\Theta$ are continuous functions of Θ by following the forward & backward update in Appx. B.3. Therefore, (1) and (2) are hold.

For (3), from Lemma 5, we have:

$$\text{minEig}(\text{Cov}(\hat{\mathbf{u}}_{\Theta^*,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_{\Theta^*}) = \text{minEig}(\mathbf{\Sigma}^*) > \eta_{\mathbf{\Sigma}^*}.$$

Because $\mathbf{A}_\Theta$ and $\hat{\mathbf{\Sigma}}_\Theta$ are continuous for Θ, for some $0 < \mathbf{s}_{\text{inf}} < \eta_{\mathbf{\Sigma}^*}$ and $\epsilon_1 = \frac{1}{c_1}(\eta_{\mathbf{\Sigma}^*} - \mathbf{s}_{\text{inf}}), \exists \tilde{r}_a$ that $\forall \Theta \in \mathbb{B}_2(\Theta^*, \tilde{r}_a)$, we have:

$$\|(\text{Cov}(\hat{\mathbf{u}}_{\Theta^*,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_{\Theta^*}) - (\text{Cov}(\hat{\mathbf{u}}_{\Theta,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_\Theta)\|_F \leq c_1\epsilon_1.$$

From Lemma 6, we have

$$|\text{minEig}(\text{Cov}(\hat{\mathbf{u}}_{\Theta,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_\Theta) - \text{minEig}(\text{Cov}(\hat{\mathbf{u}}_{\Theta^*,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_{\Theta^*})| < c_1\epsilon_1,$$

and we have:

$$\text{minEig}(\text{Cov}(\hat{\mathbf{u}}_{\Theta,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_\Theta) > \eta_{\mathbf{\Sigma}^*} - c_1\epsilon_1 > \mathbf{s}_{\text{inf}}.$$

Then (3) is hold.

For (4), $\forall j \in 1, \ldots, P$ and $l \in 1, \ldots, L_j$, we have

$$\mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta^*}(\mathbf{Y}_{jl}\hat{\mathbf{u}}^T_{jl,\mathbf{Y},\Theta^*}) = \mathbb{E}_{\mathbf{x}|\Theta^*}(\mathbf{B}^*_j\mathbf{x}_{jl}\mathbf{x}^T_{jl}) = \mathbf{B}^*_j\mathbf{\Sigma}^*_{jl}$$

where $\mathbf{B}^*_j\mathbf{\Sigma}^*_{jl}$ is column full rank since $\mathbf{B}^*_j$ is column full rank and $\mathbf{\Sigma}^*_{jl}$ is full rank. Therefore, we have $\sigma_{\min}(\mathbf{B}^*_j\mathbf{\Sigma}^*_{jl}) > 0$. Because $\mathbf{A}_\Theta$ and $\hat{\mathbf{\Sigma}}_\Theta$ are continuous to Θ, for some $0 < \mathtt{b}_{\inf} < \min_{j,l}\sigma_{\min}(\mathbf{B}^*_j\mathbf{\Sigma}^*_{jl})$ and $\epsilon_2 = \frac{1}{c_2}(\sigma_{\min}(\mathbf{B}^*_j\mathbf{\Sigma}^*_{jl}) - \mathtt{b}_{\inf}), \exists\tilde{r}_{b,jl}$ that $\forall\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r}_{b,jl})$, we have:

$$\|\mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta^*}(\mathbf{Y}_{jl}\hat{\mathbf{u}}^T_{jl,\mathbf{Y},\Theta^*}) - \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\mathbf{Y}_{jl}\hat{\mathbf{u}}^T_{jl,\mathbf{Y},\Theta})\|_F < c_2\epsilon_2.$$

From Lemma 7, we have

$$\sigma_{\min}(\mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\mathbf{Y}_{jl}\hat{\mathbf{u}}^T_{jl,\mathbf{Y},\Theta})) > \sigma_{\min}(\mathbf{B}^*_j\mathbf{\Sigma}^*_{jl}) - c_2\epsilon_2 > \mathtt{b}_{\inf} > 0.$$

Let $\tilde{r}_b = \min_{j,l}\tilde{r}_{b,jl}$, then (4) is hold.

Finally, we set $\tilde{r}_1 = \min(\tilde{r}_a, \tilde{r}_b)$ to obtain (1) to (4). □

A.4 Proof of Lemma 2

Proof. Since $G(\mathbf{C}, \Theta)$ is a continuous function of Θ, $\forall\eta_1, \eta_2, \mathbf{C}$, $\exists\tilde{r}_2$ that $\forall\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r}_2)$, we have $|G(\mathbf{C}, \Theta) - G(\mathbf{C}, \Theta^*)| < \frac{1}{2}(\eta_1 - \eta_2)$, for some $0 < \eta_2 < \eta_1$.

And from Condition 4, $\forall\pi \notin \Pi^*_0$, we have

$$G(\mathbf{C}^*, \Theta^*) - G(\mathbf{C}^*_{\Theta^*}(\pi), \Theta^*) < -\eta_1. \tag{28}$$

Then $\forall\pi \notin \Pi^*_0$, we have

$$\begin{aligned}
G(\mathbf{C}^*, \Theta) - G(\mathbf{C}^*_{\Theta}(\pi), \Theta) &\leq |G(\mathbf{C}^*, \Theta) - G(\mathbf{C}^*, \Theta^*)| \\
&\quad + G(\mathbf{C}^*, \Theta^*) - G(\mathbf{C}^*_{\Theta^*}(\pi), \Theta^*) \\
&\quad + |G(\mathbf{C}^*_{\Theta}(\pi), \Theta) - G(\mathbf{C}^*_{\Theta}(\pi), \Theta^*)| \\
&< \frac{1}{2}(\eta_1 - \eta_2) - \eta_1 - \frac{1}{2}(\eta_1 - \eta_2) \\
&= -\eta_2.
\end{aligned}$$

Therefore, $\forall \pi \notin \Pi_0^*$, we have:

$$G(\mathbf{C}_\Theta^*, \Theta) - G(\mathbf{C}_\Theta^*(\pi), \Theta) \leq G(\mathbf{C}^*, \Theta) - G(\mathbf{C}_\Theta^*(\pi), \Theta) < -\eta_2.$$

This shows that $\mathbf{C}_\Theta^*(\pi)$ is not the minimum solution of $G(\mathbf{C}, \Theta)$, and we simultaneously obtain (1) and (2). □

A.5 Proof of Lemma 3

For Lemma 3 (1):

Proof. For a fixed $\Theta \in \mathbb{B}_2(\Theta^*, \tilde{r})$, let $\hat{\mathbf{C}}$ be the estimator that minimizes $G_n(\mathbf{C}, \Theta) + \lambda\|\mathbf{C}\|_{l_1/F}$ and is consistent with causal order $\hat{\pi}$. We have

$$\begin{aligned}
&\frac{1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\mathbf{X}\mathbf{C}_\Theta^*(\hat{\pi}) - \mathbf{X}\hat{\mathbf{C}}\|_F^2 + \lambda\|\hat{\mathbf{C}}\|_{l_1/F} \\
\leq\ &\frac{1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}(\|\mathbf{X} - \mathbf{X}\mathbf{C}_\Theta^*\|_F^2 - \|\mathbf{X} - \mathbf{X}\mathbf{C}_\Theta^*(\hat{\pi})\|_F^2) \\
&\quad + \frac{2}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\langle \mathbf{X} - \mathbf{X}\mathbf{C}_\Theta^*(\hat{\pi}), \mathbf{X}(\hat{\mathbf{C}} - \mathbf{C}_\Theta^*(\hat{\pi}))\rangle + \lambda\|\mathbf{C}_\Theta^*(\hat{\pi})\|_{l_1/F} \\
\leq\ &(I) + (II) + \lambda\|\mathbf{C}_\Theta^*(\hat{\pi})\|_{l_1/F},
\end{aligned} \tag{29}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, and $\|\cdot\|_F$ denotes the Frobenius norm. Next, we will gives the upper bound for terms (I) and (II).

Bound (I):

$$\begin{aligned}
(I) &= G_n(\mathbf{C}_\Theta^*, \Theta) - G_n(\mathbf{C}_\Theta^*(\hat{\pi}), \Theta) \\
&\leq |G_n(\mathbf{C}_\Theta^*, \Theta) - G(\mathbf{C}_\Theta^*, \Theta)| + G(\mathbf{C}_\Theta^*, \Theta) - G(\mathbf{C}_\Theta^*(\hat{\pi}), \Theta) + |G_n(\mathbf{C}_\Theta^*(\hat{\pi}), \Theta) - G(\mathbf{C}_\Theta^*(\hat{\pi}), \Theta)|.
\end{aligned}$$

We have the following statements, which show that the term $G_n(\mathbf{C}, \Theta) - G(\mathbf{C}, \Theta)$ has expectation 0 and bounded variance:

(1) $\mathbb{E}_{\mathbf{Y}}(G_n(\mathbf{C}, \Theta) - G(\mathbf{C}, \Theta)) = 0$;

(2) $\mathrm{Var}(G_n(\mathbf{C}, \Theta) - G(\mathbf{C}, \Theta)) = \frac{1}{N}\mathrm{Var}(\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}\|\mathbf{x} - \mathbf{x}\mathbf{C}\|_F^2) \leq \frac{\|\mathbf{I}-\mathbf{C}\|_F^4 \mathbf{x}_{\sup}^4}{N}$.

By Chebyshev's inequality, we have:

$$P\left(|G_n(\mathbf{C},\Theta) - G(\mathbf{C},\Theta)| > \sqrt{\frac{\|\mathbf{I}-\mathbf{C}\|_F^4 \mathbf{x}_{\sup}^4}{\varrho_1 N}}\right) < \varrho_1 \tag{30}$$

Using Eq. (30) in (I), we obtain the following inequality with probability at least $1 - 2\varrho_1$:

$$(I) \leq G(\mathbf{C}_\Theta^*, \Theta) - G(\mathbf{C}_\Theta^*(\hat{\pi}), \Theta) + \sqrt{\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*\|_F^4 \mathbf{x}_{\sup}^4}{\varrho_1 N}} + \sqrt{\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*(\hat{\pi})\|_F^4 \mathbf{x}_{\sup}^4}{\varrho_1 N}} \tag{31}$$

Bound (II):

To bound the second term, we aim to show that the following equation holds true with high probability for $\delta_1 \in (0, 1/2)$:

$$\begin{aligned} &\frac{1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\langle \mathbf{X} - \mathbf{X}\mathbf{C}_\Theta^*(\hat{\pi}), \mathbf{X}(\hat{\mathbf{C}} - \mathbf{C}_\Theta^*(\hat{\pi}))\rangle \\ \leq &\frac{\delta_1}{2N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\mathbf{X}(\hat{\mathbf{C}} - \mathbf{C}_\Theta^*(\hat{\pi}))\|_F^2 + \delta_1\lambda\|\hat{\mathbf{C}} - \mathbf{C}_\Theta^*(\hat{\pi})\|_{l_1/F} \end{aligned} \tag{32}$$

Let $\mathbf{e}_j(\pi) \in \mathbb{R}^N$ as the j-th column of matrix $\mathbf{X} - \mathbf{X}\mathbf{C}_\Theta^*(\hat{\pi})$ and $\boldsymbol{\beta} \in \mathbb{R}^M$ as the j-th column of matrix $\hat{\mathbf{C}} - \mathbf{C}_\Theta^*(\hat{\pi})$. Denote $\mathcal{E}_j$ is the event:

$$\mathcal{E}_j := \left\{ \sup_{\boldsymbol{\beta}\in\mathbb{R}^M} \frac{1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\langle \mathbf{e}_j(\hat{\pi}), \mathbf{X}\boldsymbol{\beta}\rangle - \frac{\delta_1}{2N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\mathbf{X}\boldsymbol{\beta}\|_2^2 - \delta_1\lambda\|\boldsymbol{\beta}\|_{l_1/l_2} \leq 0 \right\}, \tag{33}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_P]$ for $\boldsymbol{\beta}_i \in \mathbb{R}^{L_i K_i}$ and $\|\boldsymbol{\beta}\|_{l_1/l_2} = \sum_i \|\boldsymbol{\beta}_i\|_2$.

Therefore, to prove Eq. (32), it suffices to show that for any given column j and causal order $\hat{\pi}$, the event $\mathcal{E}$ hold with a high probability.

We can then express $\mathcal{E}_j$ as:

$$\begin{aligned} \mathcal{E}_j &\subseteq \left\{ \sup_{\boldsymbol{\beta}\in\mathbb{R}^M} \frac{1}{2N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\frac{\mathbf{e}_j(\hat{\pi})}{\delta_1}\|_2^2 - \frac{1}{2N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\frac{\mathbf{e}_j(\hat{\pi})}{\delta_1} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{l_1/l_2} \leq 0 \right\} \\ &= \left\{ \mathbf{0} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^M} \frac{1}{2N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\frac{\mathbf{e}_j(\hat{\pi})}{\delta_1} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{l_1/l_2} \right\} \end{aligned}$$

Event $\mathcal{E}_j$ is correspond to the Null-consistency of group lasso problem, we use Lemma

8 to find the solution $\boldsymbol{\beta}$ and $\mathbf{w}$,

$$\boldsymbol{\beta} = \mathbf{0},$$
$$\mathbf{w} = \frac{1}{\lambda N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}(\mathbf{X}^T\frac{\mathbf{e}_j(\hat{\pi})}{\delta_1}).$$

Next we proof that $\|\mathbf{w}\|_{l_\infty/l_2} \leq 1$ holds with a high probability, where $\|\mathbf{w}\|_{l_\infty/l_2} = \max_{i=1,\ldots,P}\|\mathbf{w}_i\|_2$ and $\mathbf{w}_i$ is the gradient corresponds to $\boldsymbol{\beta}_i$. To proof this, we bound the variance of $\|\mathbf{w}\|_2$.

We first prove that the expectation of $\mathbf{w}$ is $\mathbf{0}$ from Lemma 9, and we have

$$\begin{aligned}
\|\mathbf{w}\|_2^2 &\leq \frac{1}{\lambda^2 N^2 \delta_1^2}\mathbb{E}_{\mathbf{X}|\mathcal{Y},\Theta}\|\sum_{n=1}^{N}\mathbf{x}^{(n)T}\mathbf{e}_j^{(n)}(\hat{\pi})\|_2^2 \\
&= \frac{1}{\lambda^2 N^2 \delta_1^2}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{x}^{(n)T}\mathbf{e}_j^{(n)}(\hat{\pi})\|_2^2 \\
&\leq \frac{1}{\lambda^2 N^2 \delta_1^2}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{x}^{(n)T}\mathbf{x}^{(n)}(\mathbf{I}-\mathbf{C}_\Theta^*(\pi))\|_F^2 \\
&\leq \frac{\|\mathbf{I}-\mathbf{C}_\Theta^*(\pi)\|_F^2}{\lambda^2 N^2 \delta_1^2}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{x}^{(n)}\|_2^4,
\end{aligned}$$

where,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}}\left(\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{x}^{(n)}\|_2^4\right) &\leq N\mathbf{x}_{\sup}^4, \\
\mathrm{Var}\left(\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{x}^{(n)}\|_2^4\right) &\leq N\mathbf{x}_{\sup}^8.
\end{aligned}$$

Suppose we have $\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*(\pi)\|_F^2\mathbf{x}_{\sup}^4}{\lambda^2 N\delta_1^2} < 1$. By Chebyshev's inequality, we have

$$P(\|\mathbf{w}\|_2^2 \geq 1) \leq \frac{\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*(\pi)\|_F^4}{\lambda^4 N^3\delta_1^4}\mathbf{X}_{\sup}^8}{\left(1-\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*(\pi)\|_F^2\mathbf{x}_{\sup}^4}{\lambda^2 N\delta_1^2}\right)^2} \leq \frac{\frac{\mathtt{d}_{\sup}^4}{\lambda^4 N^3\delta_1^4}\mathbf{X}_{\sup}^8}{\left(1-\frac{\mathtt{d}_{\sup}^2\mathbf{x}_{\sup}^4}{\lambda^2 N\delta_1^2}\right)^2} := \varrho_2$$

Since $\|\mathbf{w}\|_{l_\infty/l_2} \leq \|\mathbf{w}\|_2$, we have:

$$P(\|\mathbf{w}\|_{l_\infty/l_2} \geq 1) \leq \varrho_2.$$

Thus, with probability $1-\varrho_2$, event $\mathcal{E}_j$ holds true. Taking uniform control over all possible $j=1,2,\ldots,M$ and $\hat{\pi}$, we conclude that with probability $1-MP!\varrho_2$, Eq. (32) holds true.

Finally, for Lemma 3(1), suppose $\hat{\pi}\notin\Pi_0^*$, then $G(\mathbf{C}_\Theta^*,\Theta)-G(\mathbf{C}_\Theta^*(\hat{\pi}),\Theta)<-\eta_2$, and we back to Eq. (29). With probability $1-\varrho_1-MP!\varrho_2$, we have:

$$\begin{aligned}
&\frac{1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\mathbf{X}\mathbf{C}_\Theta^*(\hat{\pi})-\mathbf{X}\hat{\mathbf{C}}\|_F^2+\lambda\|\hat{\mathbf{C}}\|_{l_1/F}\\
\leq&-\eta_2+\sqrt{\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*\|_F^4\mathtt{x}_{\sup}^4}{\varrho_1 N}}+\sqrt{\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*(\hat{\pi})\|_F^4\mathtt{x}_{\sup}^4}{\varrho_1 N}}\\
&+\frac{\delta_1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\mathbf{X}(\hat{\mathbf{C}}-\mathbf{C}_\Theta^*(\hat{\pi}))\|_F^2+2\delta_1\lambda\|\hat{\mathbf{C}}-\mathbf{C}_\Theta^*(\hat{\pi})\|_{l_1/F}+\lambda\|\mathbf{C}_\Theta^*(\hat{\pi})\|_{l_1/F}.
\end{aligned}\tag{34}$$

For $\delta_1\in(0,1)$, we have:

$$\begin{aligned}
&\frac{1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\mathbf{X}\mathbf{C}_\Theta^*(\hat{\pi})-\mathbf{X}\hat{\mathbf{C}}\|_F^2\\
\leq&-\eta_2+\sqrt{\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*\|_F^4\mathtt{x}_{\sup}^4}{\varrho_1 N}}+\sqrt{\frac{\|\mathbf{I}-\mathbf{C}_\Theta^*(\hat{\pi})\|_F^4\mathtt{x}_{\sup}^4}{\varrho_1 N}}+\lambda(2\delta_1+1)\mathtt{c}_{\sup}.
\end{aligned}$$

It contradicts with the condition that:

$$\eta_2>2\sqrt{\frac{\mathtt{d}_{\sup}^4\mathtt{x}_{\sup}^4}{\varrho_1 N}}+\lambda(2\delta_1+1)\mathtt{c}_{\sup}.$$

For Lemma 3(2): we denote that $\boldsymbol{\Delta}=\hat{\mathbf{C}}-\mathbf{C}_\Theta^*(\hat{\pi})$, we have:

$$\begin{aligned}
\frac{1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\mathbf{X}\boldsymbol{\Delta}\|_F^2&=\frac{1}{N}\sum_{n=1}^N\|\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}\boldsymbol{\Delta}\|_F^2+\mathrm{tr}(\boldsymbol{\Delta}^T\hat{\boldsymbol{\Sigma}}_\Theta\boldsymbol{\Delta})\\
&=\mathrm{tr}(\boldsymbol{\Delta}^T(\frac{1}{N}\sum_{n=1}^N\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}})\boldsymbol{\Delta})+\mathrm{tr}(\boldsymbol{\Delta}^T\hat{\boldsymbol{\Sigma}}_\Theta\boldsymbol{\Delta})\\
&\geq\|\boldsymbol{\Delta}\|_F^2\mathrm{minEig}\left(\frac{1}{N}\sum_{n=1}^N\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}+\hat{\boldsymbol{\Sigma}}_\Theta\right).
\end{aligned}$$

Denote $\boldsymbol{\Phi}_\Theta:=\frac{1}{N}\sum_{n=1}^N\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}+\hat{\boldsymbol{\Sigma}}_\Theta$ and denote $\bar{\boldsymbol{\Phi}}_\Theta:=\mathbb{E}_{\mathcal{Y}}(\boldsymbol{\Phi}_\Theta)$. From Lemma

1, we have $\text{minEig}(\bar{\mathbf{\Phi}}_\Theta) > \mathtt{s}_{\text{inf}}$, and

$$\begin{aligned}\|\mathbf{\Phi}_\Theta - \bar{\mathbf{\Phi}}_\Theta\|_F^2 &= \|\frac{1}{N}\sum_{n=1}^{N}\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}} - \bar{\mathbf{\Phi}}_\Theta\|_F^2 \\ &= \frac{1}{N^2}\sum_{n=1}^{N}\|\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}} - \bar{\mathbf{\Phi}}_\Theta\|_F^2 \\ &= \frac{1}{N^2}\sum_{n=1}^{N}\|\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}} - \sum_{n'=1}^{N}(\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n')}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n')}})\|_F^2.\end{aligned}$$

where,

$$\begin{aligned}\mathbb{E}\left(\sum_{n=1}^{N}\|\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}} - \sum_{n'=1}^{N}(\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n')}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n')}})\|_F^2\right) &\leq N\mathtt{x}_{\text{sup}}^4, \\ \text{Var}\left(\sum_{n=1}^{N}\|\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n)}} - \sum_{n'=1}^{N}(\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n')}}^T\hat{\mathbf{u}}_{\Theta,\mathbf{Y}^{(n')}})\|_F^2\right) &\leq N\mathtt{x}_{\text{sup}}^8.\end{aligned}$$

By Chebyshev's inequality, we have:

$$\begin{aligned}P\left(\|\mathbf{\Phi}_\Theta - \bar{\mathbf{\Phi}}_\Theta\|_F^2 \geq \frac{\mathtt{x}_{\text{sup}}^4}{N} + \sqrt{\frac{\mathtt{x}_{\text{sup}}^8}{\varrho_3 N}}\right) &< \varrho_3, \\ P\left(|\text{minEig}(\mathbf{\Phi}_\Theta) - \text{minEig}(\bar{\mathbf{\Phi}}_\Theta)| \geq \sqrt{\frac{\mathtt{x}_{\text{sup}}^4}{N} + \sqrt{\frac{\mathtt{x}_{\text{sup}}^8}{\varrho_3 N}}}\right) &< \varrho_3, \\ P\left(\text{minEig}(\mathbf{\Phi}_\Theta) \geq \mathtt{s}_{\text{inf}} - \sqrt{\frac{\mathtt{x}_{\text{sup}}^4}{N} + \sqrt{\frac{\mathtt{x}_{\text{sup}}^8}{\varrho_3 N}}}\right) &> 1 - \varrho_3.\end{aligned}$$

Then, with at least probability $1 - 2\varrho_1 - P!M\varrho_2 - \varrho_3$, we have $\hat{\pi} \in \Pi_0^*$, therefore:

$$\begin{aligned}\|\mathbf{\Delta}\|_F^2 &\leq \frac{\frac{1}{N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\mathbf{X}\mathbf{\Delta}\|_F^2}{\mathtt{s}_{\text{inf}} - \sqrt{\frac{\mathtt{x}_{\text{sup}}^4}{N} + \sqrt{\frac{\mathtt{x}_{\text{sup}}^8}{\varrho_3 N}}}} \\ &\leq \frac{2\sqrt{\frac{\mathtt{d}_{\text{sup}}^4\mathtt{x}_{\text{sup}}^4}{\varrho_1 N}} + \lambda(2\delta_1 + 1)\mathtt{c}_{\text{sup}}}{\mathtt{s}_{\text{inf}} - \sqrt{\frac{\mathtt{x}_{\text{sup}}^4}{N} + \sqrt{\frac{\mathtt{x}_{\text{sup}}^8}{\varrho_3 N}}}}.\end{aligned}$$

□

A.6 Proof of Lemma 4

Proof. Because $\hat{\mathbf{B}}_j^T\hat{\mathbf{B}}_j = \mathbf{I}$, $\mathrm{tr}(\hat{\mathbf{B}}_j^T\hat{\mathbf{\Sigma}}_{jl}\hat{\mathbf{B}}_j) = \mathrm{tr}(\hat{\mathbf{\Sigma}}_{jl})$ is a constant unrelated to $\hat{\mathbf{B}}_j$. For a fixed j, the estimator of $\hat{\mathbf{B}}_j$ is given by:

$$\hat{\mathbf{B}}_j = \arg\min_{\mathbf{B}_j} \frac{1}{NL_j}\sum_{n=1}^{N}\sum_{l=1}^{L_j}\|\mathbf{Y}_{jl}^{(n)}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}^{(n)}}^{(n)T} - \mathbf{B}_j\|_F^2$$

$$\text{s.t.} \quad \mathbf{B}_j^T\mathbf{B}_j = \mathbf{I}.$$

We denote $\mathbf{Z} = \frac{1}{NL_j}\sum_{n=1}^{N}\sum_{l=1}^{L_j}\mathbf{Y}_{jl}^{(n)}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}^{(n)}}^{(n)T}$ and $\bar{\mathbf{Z}} = \mathbb{E}_{\mathbf{Y}}(\mathbf{Z})$. We consider $\mathbf{Z}$ is a small perturbation of $\mathbf{Z} = \bar{\mathbf{Z}} + \mathbf{E}$ and use the perturbation theory of Polar decomposition. From Li (1993), we obtain that:

$$\|\hat{\mathbf{B}}_j - \mathbf{B}_{\Theta j}^*\|_F \leq \frac{\|\mathbf{Z} - \bar{\mathbf{Z}}\|_F}{\min\{\|\mathbf{Z}^+\|_2^{-1}, \|\bar{\mathbf{Z}}^+\|_2^{-1}\}}, \tag{35}$$

where $\|\mathbf{Z}^+\|_2^{-1}$ and $\|\bar{\mathbf{Z}}^+\|_2^{-1}$ is smallest singular value of $\mathbf{Z}$ and $\bar{\mathbf{Z}}$ greater than 0. Next, we bound the numerator and denominator of RHS of Eq. (35).

For the numerator, we have

$$\begin{aligned}\|\mathbf{E}\|_F^2 &= \|\frac{1}{NL_j}\sum_{n=1}^{N}\sum_{l=1}^{L_j}\mathbf{Y}_{jl}^{(n)}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}^{(n)}}^{(n)T} - \bar{\mathbf{Z}}\|_F^2 \\ &= \frac{1}{N^2}\sum_{n=1}^{N}\|\frac{1}{L_j}\sum_{l=1}^{L_j}(\mathbf{Y}_{jl}^{(n)}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}^{(n)}}^{(n)T}) - \mathbf{Z}\|_F^2,\end{aligned}$$

where

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{Y}}\left(\sum_{n=1}^{N}\|\frac{1}{L_j}\sum_{l=1}^{L_j}(\mathbf{Y}_{jl}^{(n)}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}^{(n)}}^{(n)T})-\mathbf{Z}\|_F^2\right)\\
=&N\mathbb{E}_{\mathbf{Y}}\left(\|\frac{1}{L_j}\sum_{l=1}^{L_j}(\mathbf{Y}_{jl}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}}^{T})-\mathbf{Z}\|_F^2\right)\\
=&\frac{N}{L_j^2}\mathbb{E}_{\mathbf{Y}}\left(\|\sum_{l=1}^{L_j}(\mathbf{Y}_{jl}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}}^{T}-\mathbf{Z})\|_F^2\right)\\
\leq&\frac{N}{L_j}\mathbb{E}_{\mathbf{Y}}\left(\sum_{l=1}^{L_j}\|\mathbf{Y}_{jl}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}}^{T}\|_F^2\right)\\
\leq&N\mathtt{y}_{\sup}^2
\end{aligned}
$$

and

$$
\mathrm{Var}\left(\sum_{n=1}^{N}\|\frac{1}{L_j}\sum_{l=1}^{L_j}(\mathbf{Y}_{jl}^{(n)}\hat{\mathbf{u}}_{jl;\Theta,\mathbf{Y}^{(n)}}^{(n)T})-\mathbf{Z}\|_F^2\right)\leq N\mathtt{y}_{\sup}^4.
$$

then by Chebyshev's inequality, we have:

$$
P\left(\|\mathbf{Z}-\bar{\mathbf{Z}}\|_F^2\geq\frac{\mathtt{y}_{\sup}^2}{N}+\sqrt{\frac{\mathtt{y}_{\sup}^4}{\varrho_4 N}}\right)<\varrho_4. \tag{36}
$$

For the denominator, from Lemma 1(3). By Chebyshev's inequality, we have:

$$
P\left(\|\mathbf{Z}^{+}\|_2^{-1}\leq\mathtt{b}_{\inf}-\sqrt{\frac{\mathtt{y}_{\sup}^2}{N}+\sqrt{\frac{\mathtt{y}_{\sup}^4}{\varrho_5 N}}}\right)<\varrho_5. \tag{37}
$$

Combine Eq. (36) and Eq. (37), at least probability $1-\varrho_4-\varrho_5$, we have:

$$
\|\hat{\mathbf{B}}_j-\mathbf{B}_{\Theta j}^{*}\|_F^2\leq\frac{\frac{\mathtt{y}_{\sup}^2}{N}+\sqrt{\frac{\mathtt{y}_{\sup}^4}{\varrho_4 N}}}{\left(\mathtt{b}_{\inf}-\sqrt{\frac{\mathtt{y}_{\sup}^2}{N}+\sqrt{\frac{\mathtt{y}_{\sup}^4}{\varrho_5 N}}}\right)^2}. \tag{38}
$$

Finally, we take the uniform control for all nodes $j=1,2,\ldots,P$, then with probability

$1 - P\varrho_4 - P\varrho_5$, we have:

$$\|\hat{\mathbf{B}} - \mathbf{B}^*_\Theta\|_F^2 \leq \frac{P\left(\frac{\mathtt{y}_{\sup}^2}{N} + \sqrt{\frac{\mathtt{y}_{\sup}^4}{\varrho_4 N}}\right)}{\left(\mathtt{b}_{\inf} - \sqrt{\frac{\mathtt{y}_{\sup}^2}{N} + \sqrt{\frac{\mathtt{y}_{\sup}^4}{\varrho_5 N}}}\right)^2}.$$

□

B Minor Lemma and Derivation

B.1 Minor Lemma

Lemma 5. $\mathbb{E}_{\mathbf{Y}}(\hat{\mathbf{u}}_{\Theta^*,\mathbf{Y}}) = \mathbf{0}$ *and* $\mathrm{Cov}(\hat{\mathbf{u}}_{\Theta^*,\mathbf{Y}}) + \hat{\mathbf{\Sigma}}_{\Theta^*} = (\mathbf{I} - \mathbf{C}^*)^{-T}\omega_0^{2*}(\mathbf{I} - \mathbf{C}^*)^{-1}$.

Proof. We have

$$\mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta^*}(\mathbf{x}) = \mathbb{E}_{\mathbf{x}|\Theta^*}(\mathbf{x}) = \mathbf{0},$$

$$\mathrm{Cov}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta^*}(\mathbf{x}) + \mathbb{E}_{\mathbf{Y}}\mathrm{Cov}_{\mathbf{x}|\mathbf{Y};\Theta^*}(\mathbf{x}) = \mathrm{Cov}_{\mathbf{x}|\Theta^*}(\mathbf{x}) = (\mathbf{I} - \mathbf{C}^*)^{-T}\omega_0^{2*}(\mathbf{I} - \mathbf{C}^*)^{-1}.$$

□

Lemma 6. *For any positive definite matrix* $\mathbf{A}$ *and* $\mathbf{B}$, $\mathrm{minEig}(\mathbf{A}) - \mathrm{minEig}(\mathbf{B}) \leq \|\mathbf{A} - \mathbf{B}\|_F$.

Proof. It is straightforward from Li (1994) that we have

$$\sqrt{\sum_i (\lambda_{A,i} - \lambda_{B,r(i)})^2} \leq \|\mathbf{A} - \mathbf{B}\|_F,$$

where λ_A and λ_B are the eigenvalues of matrix $\mathbf{A}$ and $\mathbf{B}$. □

Lemma 7. *For any matrix* $\mathbf{A}$ *and* $\mathbf{B}$*, we have* $\sigma_{\min}(\mathbf{A}) - \sigma_{\min}(\mathbf{B}) \leq \|\mathbf{A} - \mathbf{B}\|_F$

Proof. It is straightforward from Mirsky (1960) that we have:

$$\sqrt{\sum_i (\sigma_{A,i} - \sigma_{B,i})^2} \leq \|\mathbf{A} - \mathbf{B}\|_F,$$

where σ_A and σ_B are the singular values of matrix $\mathbf{A}$ and $\mathbf{B}$. □

Lemma 8. *The $\boldsymbol{\beta} = 0$ is the optimal solution of the penalized Lasso with l_1/l_2 penalty $\frac{1}{2N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\|\frac{\mathbf{e}_j(\hat{\pi})}{\delta_1} - \mathbf{X}\boldsymbol{\beta}\| + \lambda\|\boldsymbol{\beta}\|_{l_1/F}$ if the following condition is hold:*

$$\begin{aligned}\mathbf{w} &\in \partial\|\boldsymbol{\beta}\|_{l_1/F},\\ \frac{1}{2N}\mathbb{E}_{\mathbf{X}|\mathcal{Y};\Theta}\mathbf{X}^T(\frac{\mathbf{e}_j(\hat{\pi})}{\delta_1} - \mathbf{X}\boldsymbol{\beta}) + \lambda\mathbf{w} &= 0,\\ \|\mathbf{w}\|_{l_\infty/l_2} &< 1.\end{aligned}$$

Proof. It is straightforward by following Lemma 1 in Aragam et al. (2015). □

Lemma 9. *$\forall\Theta$ and π, denote $\mathbf{e}_j(\pi)$ is the j-th column of $\mathbf{X}-\mathbf{X}\mathbf{C}^*_\Theta(\pi)$. We have $\mathbb{E}_\mathcal{Y}\mathbb{E}_{\mathbf{X}|\mathcal{Y},\Theta}(\mathbf{X}^T\mathbf{e}_j(\pi)) = \mathbf{0}, \forall j$.*

Proof. Since $\mathbf{e}_j(\hat{\pi})$ is the j-th column of $\mathbf{X} - \mathbf{X}\mathbf{C}^*_\Theta(\pi)$. Therefore, $\mathbf{C}^*_\Theta$ satisfies

$$\begin{aligned}\frac{\partial G(\mathbf{C}^*_\Theta, \Theta)}{\partial \mathbf{C}} &= 0,\\ \mathbb{E}_\mathbf{Y}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\mathbf{x}(\mathbf{x}_j - \mathbf{x}\mathbf{C}^*_\Theta(\hat{\pi})_j)) &= \mathbf{0}, \forall j,\\ \mathbb{E}_\mathbf{Y}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta}(\mathbf{x}\mathbf{e}_j(\pi))) &= \mathbf{0}.\end{aligned}$$

Therefore, we have $\mathbb{E}_\mathbf{Y}\mathbb{E}_{\mathbf{x}|\mathbf{Y},\Theta}(\mathbf{X}^T\frac{\mathbf{e}_j(\pi)}{\delta_1}) = \mathbf{0}$. □

Lemma 10 (Balakrishnan et al. (2017))**.** *For radius $\tilde{r} > 0$ and pair (γ, β) satisfying $0 \leq \gamma < \beta$, suppose that the function $Q(\cdot, \Theta^*)$ is globally β-strongly concave, and the Condition 6 holds on the ball $\mathbb{B}_2(\Theta^*, \tilde{r})$. Then the EM operator is contractive over $\mathbb{B}_2(\Theta^*, \tilde{r})$, in particular with:*

$$D(M(\Theta), \Theta^*) \leq \frac{\gamma}{\beta}D(\Theta, \Theta^*)$$

Lemma 11. *Denote the $\hat{\mathbf{r}}^{2*}_{jl,\Theta}$ as the variance determined by $\mathbf{B}^*_\Theta$ and $\hat{\mathbf{r}}$ as the variance determined by $\hat{\mathbf{B}}$ from Eq. (18). Under Lemma 4, we have*

$$\|\hat{\mathbf{r}}^2 - \hat{\mathbf{r}}^{2*}_{jl,\Theta}\|_2^2 \leq \frac{M\mathbf{y}^4_{\sup}}{N\varrho_6} + O(\frac{1}{N^2})$$

with probability $1 - M\varrho_6$.

Proof. We have

$$\begin{aligned}\hat{r}_{jl}^2 &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{Y}_{jl}^{(n)}-\hat{\mathbf{B}}\mathbf{x}_{jl}^{(n)}\|_2^2\\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{Y}_{jl}^{(n)}-\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}+\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}-\hat{\mathbf{B}}\mathbf{x}_{jl}^{(n)}\|_2^2\\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)},\Theta}(\|\mathbf{Y}_{jl}^{(n)}-\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}\|_2^2+\|\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}-\hat{\mathbf{B}}\mathbf{x}_{jl}^{(n)}\|_2^2+2\langle\mathbf{Y}_{jl}^{(n)}-\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)},\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}-\hat{\mathbf{B}}\mathbf{x}_{jl}^{(n)}\rangle).\end{aligned}$$

Denote $e_r = \hat{r}_{jl}^2 - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}_{jl}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{Y}_{jl}^{(n)}-\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}\|_2^2$, we have

$$\begin{aligned}e_r &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}_{jl}^{(n)}|\mathbf{Y}^{(n)},\Theta}(\|\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}-\hat{\mathbf{B}}\mathbf{x}_{jl}^{(n)}\|_2^2+2\langle\mathbf{Y}_{jl}^{(n)}-\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)},\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}-\hat{\mathbf{B}}\mathbf{x}_{jl}^{(n)}\rangle)\\ &\leq \frac{\delta_B}{N}(\mathrm{x}_{\sup}^2+2\|\mathbf{B}_{\Theta}^{*}\|_F\mathrm{x}_{\sup}^2+2\mathrm{y}_{\sup}).\end{aligned}$$

Denote $e_r' = \hat{r}_{jl,\Theta}^{*2} - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{x}_{jl}^{(n)}|\mathbf{Y}^{(n)},\Theta}\|\mathbf{Y}_{jl}^{(n)}-\hat{\mathbf{B}}_{\Theta}^{*}\mathbf{x}_{jl}^{(n)}\|_2^2$, we have

$$\mathbb{E}(e_r') = 0,$$

$$\mathrm{Var}(e_r') \leq \frac{1}{N}\mathrm{y}_{\sup}^4.$$

Therefore, using Chebyshev's inequality, with probability ϱ_6, we have

$$P\left(e_r' > \sqrt{\frac{\mathrm{y}_{\sup}^4}{N\varrho_6}}\right) < \varrho_6.$$

and with $1-\varrho_6$, we have

$$|e_r| + |e_r'| \leq \frac{\delta_B}{N}(\mathrm{x}_{\sup}^2+2\|\mathbf{B}_{\Theta}^{*}\|_F\mathrm{x}_{\sup}^2+2\mathrm{y}_{\sup}) + \sqrt{\frac{\mathrm{y}_{\sup}^4}{N\varrho_6}}.$$

Thus,

$$(\hat{r}_{jl}^2 - \hat{r}_{jl,\Theta}^{*2})^2 \leq \frac{\mathrm{y}_{\sup}^4}{N\varrho_6} + O(\frac{1}{N^2})$$

Taking uniform control of all j, l that, with probability $1 - M\varrho_6$ we have

$$\|\hat{\mathbf{r}}^2 - \hat{\mathbf{r}}_{jl,\Theta}^{2*}\|_2^2 \leq \frac{M\mathrm{y}_{\sup}^4}{N\varrho_6} + O(\frac{1}{N^2})$$

□

Lemma 12. *Denote the $\hat{\omega}_0^{*2}$ is the variance determined by $\mathbf{C}_\Theta^*$ and $\hat{\omega}_0^2$ is the variance determined by $\hat{\mathbf{C}}$ from Eq. (21). Under Lemma 3, we have*

$$\hat{\omega}_0^{*2} - \hat{\omega}_0^2 \leq \sqrt{\frac{\mathsf{d}_{\sup}^4 \mathsf{x}_{\sup}^4}{\varrho_1 N}} + \lambda(2\delta_1 + 1)\mathsf{c}_{\sup}$$

with probability 1.

Proof. From Eq. (21), we have

$$\begin{aligned}\omega_0^2 &= \frac{1}{NM}\sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}^{(n)}|\mathbf{Y}^{(n)}} \|\mathbf{x}^{(n)} - \mathbf{x}^{(n)}\hat{\mathbf{C}}\|_2^2 \\ &\leq 2\sqrt{\frac{\mathsf{d}_{\sup}^4 \mathsf{x}_{\sup}^4}{\varrho_1 N}} + \lambda(2\delta_1 + 1)\mathsf{c}_{\sup}\end{aligned}$$

□

B.2 Conditions to ensure the convergence of EM algorithm

To utilize the theorem proposed by Wang et al. (2015) and Balakrishnan et al. (2017), we denote Q as the population analog of Q_n. Condition 5 and 6 are common conditions to satisfy the convergence of EM algorithm.

$$\begin{aligned}Q(\Theta;\Theta') &= \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{x}|\mathbf{Y};\Theta'} \log f(\mathbf{x},\mathbf{Y};\Theta) \\ &= \int p(\mathbf{Y};\Theta^*) \int p(\mathbf{x}|\mathbf{Y};\Theta') \log f(\mathbf{x},\mathbf{Y};\Theta) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{Y}.\end{aligned}$$

Condition 5 (**Concavity-Smoothness**). *For any $\Theta_1, \Theta_2 \in \mathbb{B}_2(\Theta^*, \tilde{r})$, $Q(\cdot\ ;\Theta^*)$ is α-smooth, i.e., denote the θ_1, θ_2 are the vector form of parameter set Θ_1, Θ_2, we have*

$$Q(\Theta_1,\Theta^*) \geq Q(\Theta_2,\Theta^*) + (\theta_1 - \theta_2)^T \nabla Q(\Theta_2;\Theta^*) - \frac{\alpha}{2}\|\theta_2 - \theta_1\|_2,$$

and β-strongly concave, i.e.,

$$Q(\Theta_1,\Theta^*) \leq Q(\Theta_2,\Theta^*) + (\theta_1 - \theta_2)^T \nabla Q(\Theta_2;\Theta^*) - \frac{\beta}{2}\|\theta_2 - \theta_1\|_2.$$

Condition 6 (**Lipschitz-Gradient**). *For the true parameter* Θ^* *and any* $\Theta \in \mathbb{B}_2(\Theta^*, r)$*, denote* θ, θ^* *are the vector form of parameter set* Θ, Θ^**, we have:*

$$\|\nabla Q(M(\Theta); \Theta^*) - \nabla Q(M(\Theta); \Theta)\|_2 \leq \gamma \|\theta - \theta^*\|_2 \tag{39}$$

B.3 Computing Expectation

B.3.1 Forward filtering

When using forward filtering in DAG, we need to know source of the noise, this process is implement by the matrix $\mathbf{G}$ and $\mathbf{H}$, which record the coefficients of the noise from Eq. (7) and Eq. (1).

We, denote:

- $\mathbf{X} : \mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_P]$ with size $N \times \sum L_j K_j$, which is the distribution of $\mathbf{X}$ before forward filtering.
- $\tilde{\mathbf{X}} : \tilde{\mathbf{X}} = [\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_P]$ with size $N \times \sum L_j K_j$, which is the distribution of $\mathbf{X}$ after forward filtering.
- $\hat{\mathbf{X}} : \hat{\mathbf{X}} = [\hat{\mathbf{X}}_1, \ldots, \hat{\mathbf{X}}_P]$ with size $N \times \sum L_j K_j$, which is the distribution of $\mathbf{X}$ after backward smoothing.
- $\boldsymbol{\xi} : \boldsymbol{\xi} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_P]$ with size $N \times \sum L_j K_j$, which is the noise from Eq. (7).
- $\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} = [\varepsilon_{11}(t_1), \varepsilon_{11}(t_2), ..., \varepsilon_{PL_p}(t_T)]$ with size $N \times \sum L_j T$, which is the noise from Eq. (1).
- $\mathbf{G}$: Coefficient of noise (from Eq. (7)) with size $\sum L_j K_j \times \sum L_j K_j$.
- $\mathbf{H}$: Coefficient of noise (from Eq. (1)) with size $\sum L_j K_j \times \sum L_j T$.

- $\tilde{\mathbf{G}}$: Posterior coefficient of noise (from Eq. (7)) with size $\sum L_j K_j \times \sum L_j K_j$.
- $\tilde{\mathbf{H}}$: Posterior coefficient of noise (from Eq. (1)) with size $\sum L_j K_j \times \sum L_j T$.
- $\hat{\mathbf{G}}$: Coefficient of noise after backward smoothing (from Eq. (7)), with size $\sum L_j K_j \times \sum L_j K_j$.
- $\hat{\mathbf{H}}$: Coefficient of noise after backward smoothing (from Eq. (1)) with size $\sum L_j K_j \times \sum L_j T$.

Then $\mathbf{X}$, $\tilde{\mathbf{X}}$, $\hat{\mathbf{X}}$ have following representation:

$$\mathbf{x}^{(n)} = \mathbf{u}^{(n)} + \mathbf{G}\boldsymbol{\xi}^{(n)} + \mathbf{H}\boldsymbol{\varepsilon}^{(n)}$$

$$\tilde{\mathbf{x}}^{(n)} = \tilde{\mathbf{u}}^{(n)} + \tilde{\mathbf{G}}\boldsymbol{\xi}^{(n)} + \tilde{\mathbf{H}}\boldsymbol{\varepsilon}^{(n)}$$

$$\hat{\mathbf{x}}^{(n)} = \hat{\mathbf{u}}^{(n)} + \hat{\mathbf{G}}\boldsymbol{\xi}^{(n)} + \hat{\mathbf{H}}\boldsymbol{\varepsilon}^{(n)}$$

where $\hat{\mathbf{u}}$, $\tilde{\mathbf{u}}$ and $\hat{\mathbf{u}}$ represent the mean of $\mathbf{x}$, $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$.

Update for prior:

$$\begin{aligned}
\mathbf{x}_j^{(n)} &= \tilde{\mathbf{x}}^{(n)}\mathbf{C}_j + \boldsymbol{\varepsilon}_j^{(n)} \\
&= \sum_{k \in pa_j} \mathbf{C}_{kj}^T \tilde{\mathbf{u}}_k + \sum_{k \in pa_j} \mathbf{C}_{kj}^T \tilde{\mathbf{G}}_k \boldsymbol{\xi}^{(n)} \\
&\quad + \sum_{k \in pa_j} \mathbf{C}_{kj}^T \tilde{\mathbf{H}}_k \boldsymbol{\varepsilon}^{(n)} + \boldsymbol{\varepsilon}_j^{(n)}
\end{aligned}$$

Therefore, $\mathbf{x}_j^{(n)} \sim \mathcal{N}(\mathbf{u}_j^{(n)}, \mathbf{\Sigma}_j)$, where:

$$\begin{aligned}
\mathbf{u}_j^{(n)} &= \sum_{k \in pa_j} \mathbf{C}_{kj}^T \tilde{\mathbf{u}}_k^{(n)} \\
\mathbf{G}_j &= \sum_{k \in pa_j} \mathbf{C}_{kj}^T \tilde{\mathbf{G}}_k + \mathbf{I}_{\mathbf{G}}(j) \\
\mathbf{H}_j &= \sum_{k \in pa_j} \mathbf{C}_{kj}^T \tilde{\mathbf{H}}_k \\
\mathbf{\Sigma}_j &= \omega_0^2 \mathbf{G}_j \mathbf{G}_j^T + \mathbf{H}_j \text{diag}(\mathbf{r}) \mathbf{H}_j^T
\end{aligned}$$

where $\mathbf{I}_{\mathbf{G}}(j)$ is a $\sum L_j K_j \times \sum L_j K_j$ matrix with the identity matrix in the submatrix corresponding to node j, $\mathbf{I}_{\mathbf{G}}(j)_{jj} = \mathbf{I}_{L_j K_j \times L_j K_j}$.

Update for posterior: We estimated the posterior distribution of $\mathbf{x}$ in n-th sample,

$$\begin{aligned}
\mathbf{Y}_{jl}^{(n)} &= \mathbf{B}_j \mathbf{x}_{jl}^{(n)} + \boldsymbol{\varepsilon}_{jl}^{(n)} \\
&= \mathbf{B}_j (\mathbf{u}_{jl}^{(n)} + \mathbf{G}_{jl} \boldsymbol{\xi}^{(n)} + \mathbf{H}_{jl} \boldsymbol{\varepsilon}^{(n)}) + \boldsymbol{\varepsilon}_{jl}^{(n)}
\end{aligned}$$

Therefore, $\mathbf{Y}_{jl}^{(n)} \sim \mathcal{N}(\mathbf{B}_j \mathbf{u}_{jl}^{(n)}, \mathbf{B}_j \mathbf{\Sigma}_{jl} \mathbf{B}_j^T + r_{jl}^2 \mathbf{I}_T)$, where: And we have:

$$\begin{pmatrix} \mathbf{x}_{jl}^{(n)} \\ \mathbf{Y}_{jl}^{(n)} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \hat{\mathbf{u}}_{jl}^{(n)} \\ \mathbf{B}_j \mathbf{u}_{jl}^{(n)} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{jl} & \mathbf{\Sigma}_{jl} \mathbf{B}_j^T \\ \mathbf{B}_j \mathbf{\Sigma}_{jl} & \mathbf{B}_j \mathbf{\Sigma}_{jl} \mathbf{B}_j^T + r_{jl}^2 \mathbf{I}_T \end{pmatrix} \right)$$

The posterior $\mathbf{x}_{jl} | \mathbf{Y}_{jl} \sim \mathcal{N}(\tilde{\mathbf{u}}_{jl}, \tilde{\mathbf{\Sigma}}_{jl})$, where

$$\begin{aligned}
\tilde{\mathbf{u}}_{jl}^{(n)} &= \mathbf{u}_{jl}^{(n)} + \mathbf{\Sigma}_{jl} \mathbf{B}_j^T (\mathbf{B}_j \mathbf{\Sigma}_{jl} \mathbf{B}_j^T + r_{jl}^2 \mathbf{I}_T)^{-1} (\mathbf{Y}_{jl} - \mathbf{B}_j \mathbf{u}_{jl}^{(n)}) \\
\tilde{\mathbf{G}}_{jl} &= \mathbf{G}_{jl} - \mathbf{\Sigma}_{jl} \mathbf{B}_j^T (\mathbf{B}_j \mathbf{\Sigma}_{jl} \mathbf{B}_j^T + r_{jl}^2 \mathbf{I}_T)^{-1} \mathbf{B} \mathbf{G}_{jl} \\
\tilde{\mathbf{H}}_{jl} &= \mathbf{H}_{jl} - \mathbf{\Sigma}_{jl} \mathbf{B}_j^T (\mathbf{B}_j \mathbf{\Sigma}_{jl} \mathbf{B}_j^T + r_{jl}^2 \mathbf{I}_T)^{-1} (\mathbf{B} \mathbf{H}_{jl} + \mathbf{I}_{\mathbf{H}}(j, l))
\end{aligned}$$

where $\mathbf{I}_{\mathbf{H}}(j, l)$ is a $\sum L_j T \times \sum L_j T$ matrix with the identity matrix $\mathbf{I}_T$ in the submatrix corresponding to the l-th function in node j, $\mathbf{I}_{\mathbf{H}}(j, l)_{jl,jl} = \mathbf{I}_T$.

B.3.2 Backward smoothing

For k and the descendants j, we derive the covariance of nodes j, k:

$$\tilde{\mathbf{\Sigma}}_{j,k} = \omega_0^2 \tilde{\mathbf{G}}_j \tilde{\mathbf{G}}_k^T + \tilde{\mathbf{H}}_j D(\mathbf{r}) \tilde{\mathbf{H}}_k^T$$

$$\begin{pmatrix} \tilde{\mathbf{x}}_k^{(n)} \\ \tilde{\mathbf{x}}_{de(k)}^{(n)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \tilde{\mathbf{u}}_k^{(n)} \\ \tilde{\mathbf{u}}_{de(k)}^{(n)} \end{pmatrix}, \begin{pmatrix} \tilde{\mathbf{\Sigma}}_k & \tilde{\mathbf{\Sigma}}_{k,de(k)} \\ \tilde{\mathbf{\Sigma}}_{k,de(k)}^T & \tilde{\mathbf{\Sigma}}_{de(k)} \end{pmatrix} \right) \tag{40}$$

Derive $p(\tilde{\mathbf{x}}_k | \tilde{\mathbf{x}}_{de(k)}, \mathbf{Y})$:

$$\hat{\mathbf{u}}_k^{(n)} = \tilde{\mathbf{u}}_k^{(n)} + \tilde{\mathbf{\Sigma}}_{k,de(k)} \mathbf{\Sigma}_{de(k)}^{-1} (\hat{\mathbf{u}}_{de(k)}^{(n)} - \tilde{\mathbf{u}}_{de(k)}^{(n)})$$

$$\hat{\mathbf{G}}_k = \tilde{\mathbf{G}}_k - \tilde{\mathbf{\Sigma}}_{k,de(k)} \mathbf{\Sigma}_{de(k)}^{-1} (\tilde{\mathbf{G}}_{de(k)} - \hat{\mathbf{G}}_{de(k)})$$

$$\hat{\mathbf{H}}_k = \tilde{\mathbf{H}}_k - \tilde{\mathbf{\Sigma}}_{k,de(k)} \mathbf{\Sigma}_{de(k)}^{-1} (\tilde{\mathbf{H}}_{de(k)} - \hat{\mathbf{H}}_{de(k)})$$

Finally, posterior mean of $\mathbf{x}$ is $\hat{\mathbf{u}}$ and the posterior variance is:

$$\hat{\mathbf{\Sigma}} = \omega_0^2 \hat{\mathbf{G}} \hat{\mathbf{G}}^T + \hat{\mathbf{H}} \text{diag}(\mathbf{r}) \hat{\mathbf{H}}^T$$