Journal of Machine Learning Research 23 (2022) 1-34 Submitted 10/21; Revised 8/22; Published 9/22

Learning linear non-Gaussian directed acyclic graph with diverging number of nodes

Ruixuan Zhao RUIXUZHAO2-C@MY.CITYU.EDU.HK
School of Data Science
City University of Hong Kong
Kowloon Tong, Kowloon, Hong Kong

Xin He HE.XIN17@MAIL.SHUFE.EDU.CN
School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai, China

Junhui Wang JUNHUIWANG@CUHK.EDU.HK
Department of Statistics
The Chinese University of Hong Kong
Shatin, New Territory, Hong Kong

Editor: Garvesh Raskutti

Abstract

An acyclic model, often depicted as a directed acyclic graph (DAG), has been widely employed to represent directional causal relations among collected nodes. In this article, we propose an efficient method to learn linear non-Gaussian DAG in high dimensional cases, where the noises can be of any continuous non-Gaussian distribution. The proposed method leverages the concept of topological layer to facilitate the DAG learning, and its theoretical justification in terms of exact DAG recovery is also established under mild conditions. Particularly, we show that the topological layers can be exactly reconstructed in a bottom-up fashion, and the parent-child relations among nodes can also be consistently established. The established asymptotic DAG recovery is in sharp contrast to that of many existing learning methods assuming parental faithfulness or ordered noise variances. The advantage of the proposed method is also supported by the numerical comparison against some popular competitors in various simulated examples as well as a real application on the global spread of COVID-19.

Keywords: Causal inference, DAG, non-Gaussian noise, structural equation model, topological layer

1. Introduction

A directed acyclic graph (DAG) provides an elegant way to represent directional or causal structures among collected nodes, which finds applications in a broad variety of domains, including genetics (Sachs et al., 2005), finance (Sanford and Moosa, 2012) and social science (Newey et al., 1999). In recent years, learning the DAG structures from observed data has attracted tremendous attention from both academia and industries.

∗. Xin He and Junhui Wang are the co-corresponding authors.

In literature, various structure learning methods have been proposed to recover the Markov equivalence class (Spirtes et al., 2000; Peters et al., 2017) of a DAG, which can be roughly categorized into three classes. The first class is the constraint-based method (Spirtes et al., 2000; Kalisch and Bühlmann, 2007), which uses some local conditional independence criterion to test pairwise causal relations. The second class, referred as the score-based method (Chickering, 2002; Zheng et al., 2018; Yuan et al., 2019), attempts to optimize some goodness-of-fit measures among the possible graph space. The last class (Tsamardinos et al., 2006; Nandy et al., 2018) combines the constraint-based method and score-based method. Although success has been widely reported, most aforementioned methods can only recover the Markov equivalence class and their computational burden remains a severe bottleneck. Recently, substantial effort has been made to pursuit exact DAG recovery. For instances, Peters and Bühlmann (2014) shows that a linear Gaussian DAG is identifiable under the equal noise variance assumption, and Ghoshal and Honorio (2018) and Park (2020) relax the Gaussianity assumption but still require an explicit order among noise variances. Along this line, many learning methods are proposed to recover the exact DAG structure (Ghoshal and Honorio, 2018; Chen et al., 2019; Yuan et al., 2019; Li et al., 2020; Park, 2020), yet these assumptions of Gaussianity or ordered noise variances are often difficult to verify in practice.

Linear non-Gaussian DAG, also known as linear non-Gaussian acyclic model (LiNGAM; Shimizu et al. 2006), relaxes the Gaussianity assumption, and its identifiability does not require any additional noise variance assumption. It is clear that linear non-Gaussian DAG can accommodate more flexible distributions, and thus has attracted tremendous interests in recent years (Shimizu et al., 2006; Hoyer et al., 2008; Entner and Hoyer, 2010; Shimizu et al., 2011; Hyvarinen and Smith, 2013; Tashiro et al., 2014; Wang and Drton, 2020). Specifically, Shimizu et al. (2006) proposes an iterative search algorithm to recover the causal ordering of a linear non-Gaussian DAG by using linear independent component analysis (ICA) and permutation. Subsequently, Shimizu et al. (2011) proposes a multiple-step algorithm to learn a linear non-Gaussian DAG by some pairwise statistics, which is further extended in Hyvarinen and Smith (2013) to iteratively identify pairwise causal ordering by likelihood ratio tests. These methods often require heavy computational cost, which may lead to practical difficulties even when dealing with a medium-sized DAG. Most recently, Wang and Drton (2020) proposes a modified direct learning algorithm for a linear non-Gaussian DAG in high dimensional cases with theoretical guarantee, which sequentially recovers the causal ordering with a moment-based criterion and reconstructs the directed structure with hard-thresholding. Yet, it requires the parental faithfulness condition and its computational complexity is of exponential order of the maximum in-degree. It is worthy pointing out that some most recent methods (Ghoshal and Honorio, 2018; Park et al., 2021) can also be used to learn linear non-Gaussian DAG under high dimensional settings, yet their numerical performance and theoretical justification highly rely on the validity of the required ordered noise variance assumption.

In this paper, we propose a method to learn a linear non-Gaussian DAG with a large number of nodes, by leveraging the concept of topological layer to facilitate efficient DAG learning. It assures that any DAG can be reformulated into a unique topological structure with T layers, where the parents of a node must belong to its upper layers, and thus acyclicity is naturally guaranteed. More importantly, we show that the topological layers can be exactly reconstructed via precision matrix estimation and independence testing procedure in a bottom-up fashion, and the parent-child relations can be directly obtained from the estimated precision matrix. These results are obtained without requiring the popular faithfulness (Uhler et al., 2013; Peters et al., 2017), parental faithfulness assumption (Wang and Drton, 2020) or the ordered noise variance assumption (Ghoshal

and Honorio, 2018). The constructive proof also motivates an efficient learning algorithm for the proposed method, whose complexity is much smaller than most existing linear non-Gaussian DAG learning methods (Shimizu et al., 2006, 2011; Wang and Drton, 2020).

The main contribution of this paper is the development of an efficient method to learn linear non-Gaussian DAG in high dimensional cases, and the investigation on its statistical guarantee in terms of exact DAG recovery. More precisely, we show that the topological layers of the DAG can be exactly reconstructed in Theorem 1 and Corollary 1, and the parent-child relations can be directly recovered along with the topological layers in Corollary 2. We connect learning method and precision matrix estimation by proving that the topological layers can be exactly reconstructed via precision matrix estimations in a bottom-up fashion, and the parent-child relations can be obtained directly from the obtained precision matrix. The statistical guarantees of the proposed method by using graphical Lasso and distance covariance measure is established with sub-Gaussian and $(4m)$-th bounded moment noise distributions, respectively. The established consistency results are governed by the sample size, the number of nodes, the maximum cardinality of Markov blankets (Peters et al., 2017) and the number of topological layers. Most interestingly, the obtained results allow the number of nodes, the maximum cardinality of Markov blankets and the number of layers to diverge with the sample size at some fast rate, which is particularly attractive in a high-dimensional learning method. We want to also emphasize that the proposed method differs from most existing LiNGAM algorithms (Shimizu et al., 2006, 2011; Hyvarinen and Smith, 2013; Wang and Drton, 2020) and some very recent works on linear structural equation models (Ghoshal and Honorio, 2018; Park, 2020; Park et al., 2021), which are mostly algorithm-based (Shimizu et al., 2006, 2011; Hyvarinen and Smith, 2013), or requires the parental faithfulness assumption (Wang and Drton, 2020) or the ordered noise variance condition (Ghoshal and Honorio, 2018; Park, 2020; Park et al., 2021) for establishing DAG recovery consistency, not to mention that many of them may suffer from heavy computational cost even when dealing with a medium-sized DAG .

The rest of this paper is organized as follows. Section 2 introduces some background of linear non-Gaussian DAG. Section 3 introduces the concept of topological layers and shows that the topological layers and the parent-child relations can be exactly reconstructed in a bottom-up fashion. Section 4 provides an efficient learning algorithm for linear non-Gaussian DAG in high dimensional cases, and Section 5 establishes the reconstruction consistency of the proposed method under mild conditions. Section 6 compares the proposed method with some existing competitors in terms of their computational complexity and theoretical guarantees. Numerical experiments on several simulated examples and one real application to the spread of COVID-19 are conducted in Section 7. Section 8 contains a brief discussion, and all the technical details are provided in Appendix.

2. Preambles

Consider a DAG $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, encoding the joint distribution $P(\mathbf{x})$ of $\mathbf{x} = (x_1, ..., x_p)^T \in \mathcal{R}^p$, where $\mathcal{N} = \{1, \ldots, p\}$ consists of a set of nodes associated with each coordinate of $\mathbf{x}$, and $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ consists of all the directed edges among the nodes. The directed edge from node j to node k is denoted as $j \to k$, indicating their parent-child relationship. For simplicity, we denote node k's parents as pa_k, its children as ch_k, its descendants as de_k, its non-descendants as nd_k, and its Markov blanket as $\text{mb}_k = \text{pa}_k \cup \text{ch}_k \cup \{i \in \text{pa}_j \backslash \{k\} | j \in \text{ch}_k\}$. It is also assumed that $\mathcal{G}$ satisfies the Markov property (Spirtes et al., 2000), and thus $P(\mathbf{x})$ can be factorized as $P(\mathbf{x}) = \prod_{k=1}^{p} P(x_k | \mathbf{x}_{\text{pa}_k})$, where $\mathbf{x}_{\text{pa}_k} = \{x_j : j \in \text{pa}_k\}$.

Once each node x_k is centered with mean zero, the graph structure in $\mathcal{G}$ can be embedded into a linear structural equation model (SEM),

$$x_k = \sum_{j \in \text{pa}_k} \beta_{kj} x_j + \epsilon_k;\ k = 1, ..., p, \tag{1}$$

where $\beta_{kj} \neq 0$ for any $j \in \text{pa}_k$, ϵ_k denotes a continuous non-Gaussian noise with variance σ_k^2, and $\epsilon_l \perp\!\!\!\perp \epsilon_k$ for any $l \neq k$. This independent noise condition further implies that $\epsilon_k \perp\!\!\!\perp x_l$ for any $l \notin \text{de}_k \cup \{k\}$. It is also often assumed that there is no unobserved confounding effect among the observed nodes in $\mathcal{G}$, which is known as the casual sufficiency condition in literature (Spirtes et al., 2000).

Note that the SEM model in (1) can be organized into a matrix form

$$\mathbf{x} = \mathbf{B}\,\mathbf{x} + \boldsymbol{\epsilon},$$

where $\mathbf{B} = (\beta_{kj})_{k,j} \in \mathcal{R}^{p \times p}$ and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_p)^T$ is the noise vector with covariance matrix $\boldsymbol{\Omega} = \text{diag}\{\sigma_1^2, ..., \sigma_p^2\}$. Simple algebra yields that

$$\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\epsilon} = \mathbf{A}\boldsymbol{\epsilon}, \tag{2}$$

and $x_k = \sum_{j=1}^{p} a_{kj}\epsilon_j$, where a_{kj} is the (k, j)-th element of $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$, representing the total effect of x_j on x_k. The SEM model implies that $\epsilon_j \perp\!\!\!\perp x_k$ and thus $a_{kj} = 0$ for any node $j \in \text{de}_k$. Moreover, the covariance matrix of $\mathbf{x}$ is $\boldsymbol{\Sigma} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Omega}(\mathbf{I} - \mathbf{B})^{-T}$, and the corresponding precision matrix is $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} = (\mathbf{I} - \mathbf{B})^T\boldsymbol{\Omega}^{-1}(\mathbf{I} - \mathbf{B})$. In the sequel, we use Θ_{lk} to denote the (l, k)-th element of $\boldsymbol{\Theta}$, and $\boldsymbol{\Theta}_{-ll}$ to denote the l-th column of $\boldsymbol{\Theta}$ without Θ_{ll}.

3. Topological layers

In this section, we introduce the concept of topological layer, which allows us to convert a DAG into a unique topological structure. Particularly, given a DAG $\mathcal{G}$, we construct its topological structure by assigning each node to one and only one layer, based on its longest distance to one of the leaf nodes.

Without loss of generality, we assume $\mathcal{G}$ has a total of T layers, and $\mathcal{A}_t$ denotes all the nodes contained in the t-th layer, for $t = 0, \ldots, T-1$. It is clear that $\cup_{t=0}^{T-1}\mathcal{A}_t = \mathcal{N}$, and all the leaf nodes and isolated nodes in $\mathcal{G}$ belong to the lowest layer $\mathcal{A}_0$. For each node $k \in \mathcal{A}_t$, it follows from the layer construction that $\text{pa}_k \subset \mathcal{S}_{t+1} = \cup_{d=t+1}^{T-1}\mathcal{A}_d$, and thus acyclicity is automatically guaranteed. Note that $\mathcal{S}_0 = \mathcal{N}$. Figure 1 illustrates a toy DAG in the left panel, and its converted topological structure with three layers in the right panel.

In Figure 1, node 4 is an isolated node and node 3 has no child node, and thus they both belong to $\mathcal{A}_0$. Node 3 has two parent nodes, where node 2 belongs to $\mathcal{A}_1$ but node 1 belongs to $\mathcal{A}_2$ due to the existence of a longer path $1 \to 2 \to 3$. It is clear that the concept of topological layer is general and it can restructure any DAG in such a way that causal ordering among each layers is uniquely determined. This is in sharp contrast to the idea of causal ordering in literature (Shimizu et al., 2006, 2011; Wang and Drton, 2020), which only requires that each node is ranked behind its parents. For the toy DAG in Figure 1, it induces multiple possible causal orderings among the four nodes, such as $1 \to 2 \to 3 \to 4$, $1 \to 2 \to 4 \to 3$, $1 \to 4 \to 2 \to 3$, or $4 \to 1 \to 2 \to 3$. This

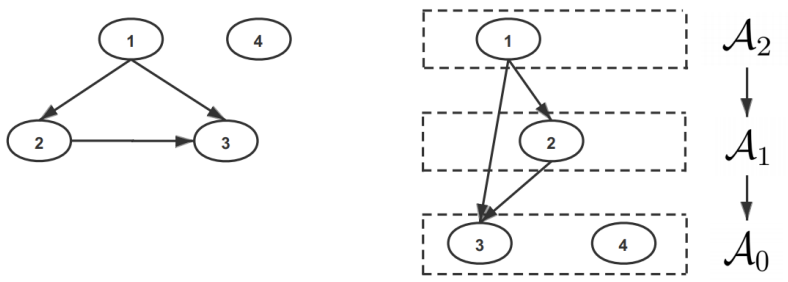


Figure 1: A toy DAG and its topological structure with three layers.

indeterministic causal ordering may cause unnecessary estimation instability and computational inefficiency in reconstructing the DAG structures.

The concept of topological layer has been considered in recent literature (Gao et al., 2020; Zhou et al., 2022). Both works define the layers in a top-down fashion based on the longest distances to the root nodes, but the topological layers in this paper is constructed in a bottom-up fashion motivated by the theoretical findings on linear non-Gaussian DAG in Section 3.1. Particularly, this bottom-up layer structure motivates us to develop a new learning method for the linear non-Gaussian DAG, which connects DAG learning with precision matrix estimation and leads to sound theoretical justification.

3.1 Reconstruction of linear non-Gaussian DAG

To be self-contained, we first restate the Darmois-Skitovitch theorem (Darmois, 1953; Skitovitch, 1953), which is crucial for the reconstruction of topological layers of a linear non-Gaussian DAG.

Lemma 1 *(Darmois-Skitovitch, 1953) Define two random variables u_1 and u_2 as linear combinations of independent random variables $s_i, i = 1, ..., m$, that $u_1 = \sum_{i=1}^{m} c_{1,i} s_i$ and $u_2 = \sum_{i=1}^{m} c_{2,i} s_i$. Then, if u_1 and u_2 are independent, all variables s_i with $c_{1,i} c_{2,i} \neq 0$ are Gaussian distributed.*

Lemma 1 shows that if s_i's are non-Gaussian distributed, it is impossible to construct two independent linear combinations of s_i's. This fact motivates us to use independence test to identify nodes in each layers in a bottom-up fashion.

Theorem 1 *Suppose that $\mathbf{x} = (x_1, .., x_p)^T \in \mathcal{R}^p$ is generated from the linear SEM model in* (1) *with precision matrix $\mathbf{\Theta}$. For any $l \in \mathcal{N}$, we regress x_l on all other nodes $\mathbf{x}_{\mathcal{N}\setminus\{l\}}$, and denote the expected residual as $e_{l,\mathcal{N}} = x_l - \mathbf{x}_{\mathcal{N}\setminus\{l\}}^T \mathbf{M}_{\mathcal{N}}^{(l)}$, where $\mathbf{M}_{\mathcal{N}}^{(l)} = -\mathbf{\Theta}_{-ll}/\Theta_{ll}$. Then, we have $l \in \mathcal{A}_0$ if and only if $e_{l,\mathcal{N}} \perp\!\!\!\perp x_k$ for any $k \in \mathcal{N}\setminus\{l\}$.*

Theorem 1 provides a sufficient and necessary condition to identify nodes in $\mathcal{A}_0$. After all the nodes in $\mathcal{A}_0$ are identified, we can remove them from $\mathcal{N}$ and denote $\mathcal{S}_1 = \mathcal{N} \setminus \mathcal{A}_0$. Next, we apply a similar treatment to $\mathcal{S}_1$ as in Theorem 1 to identify $\mathcal{A}_1$, and then $\mathcal{A}_2$, until all nodes are assigned to layers. Denote $\mathbf{\Theta}^{\mathcal{S}_t}$ as the precision matrix of $\mathbf{x}_{\mathcal{S}_t}$ and $\mathbf{\Theta}^{\mathcal{S}_0} = \mathbf{\Theta}$. Further, denote $[\Theta^{\mathcal{S}_t}]_{lk}$ as the element of $\mathbf{\Theta}^{\mathcal{S}_t}$ corresponding to nodes l and k, and $[\mathbf{\Theta}^{\mathcal{S}_t}]_{-ll}$ as the column of $\mathbf{\Theta}$ corresponding to

node l without $[\Theta^{\mathcal{S}_t}]_{ll}$. Corollary 1 summarizes the reconstruction of all the topological layers for a linear non-Gaussian DAG.

Corollary 1 *Suppose that all the conditions in Theorem 1 are satisfied and the layers $\mathcal{A}_0, ..., \mathcal{A}_{t-1}$ have been reconstructed. For any $l \in \mathcal{S}_t$, we regress x_l on $\mathbf{x}_{\mathcal{S}_t \setminus \{l\}}$ and denote the expected residual as $e_{l,\mathcal{S}_t} = x_l - \mathbf{x}_{\mathcal{S}_t \setminus \{l\}}^T \mathbf{M}_{\mathcal{S}_t}^{(l)}$ with $\mathbf{M}_{\mathcal{S}_t}^{(l)} = -[\mathbf{\Theta}^{\mathcal{S}_t}]_{-ll}/[\Theta^{\mathcal{S}_t}]_{ll}$. Then we have $l \in \mathcal{A}_t$ if and only if $e_{l,\mathcal{S}_t} \perp\!\!\!\perp x_k$ for any $k \in \mathcal{S}_t \setminus \{l\}$.*

The proof of Corollary 1 is similar to that of Theorem 1 with slight modification by replacing $\mathcal{N}$ with $\mathcal{S}_t$. The reconstruction of the topological layers follows immediately by applying Corollary 1 repeatedly in the sense of mathematical induction. Furthermore, the parent set of each node in the DAG can be determined during the reconstruction of the topological layers as well.

Corollary 2 *Suppose that all the conditions in Theorem 1 are satisfied. For any $l \in \mathcal{A}_t$ and $k \in \mathcal{S}_{t+1}$, we have $\beta_{lk} = -[\Theta^{\mathcal{S}_t}]_{lk}/[\Theta^{\mathcal{S}_t}]_{ll}$, and thus $pa_l = \{k \in \mathcal{S}_{t+1} : [\Theta^{\mathcal{S}_t}]_{lk} \neq 0\}$.*

Corollary 2 follows immediately after Corollary 1, and assures that the parent-child relations can also be sequentially reconstructed along with the topological layers. It is important to point out that both Corollaries 1 and 2 do not require the popularly-adopted faithfulness assumption (Uhler et al., 2013; Peters et al., 2017), the parental faithfulness assumption (Wang and Drton, 2020), or the ordered noise variance assumption (Ghoshal and Honorio, 2018).

3.2 An illustrative example

Consider a simple linear non-Gaussian DAG generated as follows,

$$x_1 = \epsilon_1, \;\; x_2 = \beta_{21} x_1 + \epsilon_2, \;\; x_3 = \beta_{31} x_1 + \beta_{32} x_2 + \epsilon_3, \text{ and } x_4 = \epsilon_4, \tag{3}$$

where ϵ_l is a non-Gaussian distributed noise, and $\epsilon_l \perp\!\!\!\perp \epsilon_k$ for any $l \neq k$. It is clear that the parental faithfulness assumption is violated when $\beta_{32}\beta_{21} + \beta_{31} = 0$ as illustrated in Wang and Drton (2020), and so is the faithfulness assumption. More precisely, if $\beta_{32}\beta_{21} + \beta_{31} = 0$, it follows from (3) that

$$x_3 = (\beta_{32}\beta_{21} + \beta_{31})\epsilon_1 + \beta_{32}\epsilon_2 + \epsilon_3 = 0 + \beta_{32}\epsilon_2 + \epsilon_3,$$

and thus x_3 and $x_1 = \epsilon_1$ are independent since ϵ_1, ϵ_2 and ϵ_3 are mutually independent by the generating scheme (3). However, x_3 and x_1 are not d-separated in the graph, and thus the faithfulness assumption is violated by Definition 6.33 of Peters et al. (2017). In contrast, such a linear non-Gaussian DAG can be successfully reconstructed by the proposed criteria in Section 3.1.

Next, we show that all the layers and the parent-child relations can be exactly recovered from the data by Theorem 1, Corollaries 1 and 2. We first regress x_l on $\mathbf{x}_{-l}$ and obtain the expected residuals $e_{l,\mathcal{N}}$ for $l = 1, \ldots, 4$. It is easy to see that $e_{3,\mathcal{N}} = \epsilon_3$ and $e_{4,\mathcal{N}} = \epsilon_4$, and each of them is independent with all the other three nodes. For $e_{1,\mathcal{N}}$ and $e_{2,\mathcal{N}}$, it follows from a similar treatment as in the proof of Theorem 1 that

$$\begin{aligned}
e_{1,\mathcal{N}} &= \big(1 - M_2^{(1)}\beta_{21} - M_3^{(1)}(\beta_{32}\beta_{21} + \beta_{31})\big)\epsilon_1 - (M_2^{(1)} + M_3^{(1)}\beta_{32})\epsilon_2 - M_3^{(1)}\epsilon_3, \\
e_{2,\mathcal{N}} &= \big(\beta_{21} - M_1^{(2)} - M_3^{(2)}(\beta_{32}\beta_{21} + \beta_{31})\big)\epsilon_1 + (1 - M_3^{(2)}\beta_{32})\epsilon_2 - M_3^{(2)}\epsilon_3,
\end{aligned}$$

where $M_k^{(l)}$ denotes the k-th element of $\mathbf{M}_{\mathcal{N}}^{(l)} = -\mathbf{\Theta}_{-ll}/\Theta_{ll}$ and can be regarded as the partial correlation between nodes k and l given all the other nodes in $\mathcal{N}$. Clearly, $e_{1,\mathcal{N}}$ and $e_{2,\mathcal{N}}$ are not equivalent to ϵ_1 and ϵ_2, but they are sufficient for the independent test to detect $\mathcal{A}_0$. In fact, since $\Theta_{31} = -\sigma_3^{-2}\beta_{31} \neq 0$ and $\Theta_{32} = -\sigma_3^{-2}\beta_{32} \neq 0$, $M_3^{(1)}$ and $M_3^{(2)}$ are nonzero, and thus both $e_{1,\mathcal{N}}$ and $e_{2,\mathcal{N}}$ are dependent with x_3, leading to $\mathcal{A}_0 = \{3,4\}$ and $\mathcal{S}_1 = \{1,2\}$. It also follows from Corollary 2 that $\text{pa}_3 = \{1,2\}$ and $\text{pa}_4 = \emptyset$.

Next, we repeat the above procedure for x_1 and x_2, and obtain $e_{1,\mathcal{S}_1} = (1 - M_2^{(1)}\beta_{21})\epsilon_1 - M_2^{(1)}\epsilon_2$ and $e_{2,\mathcal{S}_1} = \epsilon_2$. Clearly, $e_{2,\mathcal{S}_1}$ is independent with $x_1 = \epsilon_1$, whereas $e_{1,\mathcal{S}_1}$ is dependent with $x_2 = \beta_{21}x_1 + \epsilon_2$ due to the fact that both $e_{1,\mathcal{S}_1}$ and x_2 depend on ϵ_2 as long as $M_2^{(1)}$ is nonzero. Therefore, we have $\mathcal{A}_1 = \{2\}$ and $\text{pa}_2 = \{1\}$ by Corollaries 1 and 2. Finally, the last remaining node 1 is assigned to $\mathcal{A}_2$, and the topological layers and directed structure of the DAG in (3) are perfectly reconstructed.

4. DAG learning algorithm

Given a sample matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathcal{R}^{n\times p}$ with $\mathbf{x}_l = (x_{1l}, ..., x_{nl})^T$, we first obtain the estimated precision matrix $\widehat{\mathbf{\Theta}}$, and then the residual is $\widehat{\mathbf{e}}_{l,\mathcal{N}} = \mathbf{x}_l + \mathbf{X}_{-l}\widehat{\mathbf{\Theta}}_{-ll}/\widehat{\Theta}_{ll}$, where $\mathbf{X}_{-l}$ denotes the sample matrix without $\mathbf{x}_l$. To invoke Theorem 1, we proceed to test the independence between $\widehat{\mathbf{e}}_{l,\mathcal{N}}$ and all the nodes in $\mathcal{N} \setminus \{l\}$, and estimate $\mathcal{A}_0$ as

$$\widehat{\mathcal{A}}_0 = \Big\{ l : \widehat{\mathbf{e}}_{l,\mathcal{N}} \text{ is tested to be independent with } x_k \text{ for any } k \in \mathcal{N}\backslash\{l\} \Big\}.$$

Let $\widehat{\mathcal{S}}_1 = \mathcal{N}\backslash\widehat{\mathcal{A}}_0$, then Corollary 2 implies that for each $l \in \widehat{\mathcal{A}}_0$,

$$\widehat{\text{pa}}_l = \Big\{ k \in \widehat{\mathcal{S}}_1 : \widehat{\Theta}_{lk} \neq 0 \Big\},$$

and $\widehat{\beta}_{lk} = -\widehat{\Theta}_{lk}/\widehat{\Theta}_{ll}$ for any $k \in \widehat{\text{pa}}_l$.

Suppose that the estimated layers $\widehat{\mathcal{A}}_0, ..., \widehat{\mathcal{A}}_{t-1}$ are obtained, we denote $\widehat{\mathcal{S}}_t = \mathcal{N}\backslash\{\cup_{d=0}^{t-1}\widehat{\mathcal{A}}_d\}$ and estimate the corresponding precision matrix $\widehat{\mathbf{\Theta}}^{\widehat{\mathcal{S}}_t}$. The residuals can be computed as $\widehat{\mathbf{e}}_{l,\widehat{\mathcal{S}}_t} = \mathbf{x}_l + \mathbf{X}_{\widehat{\mathcal{S}}_t\backslash\{l\}}[\widehat{\mathbf{\Theta}}^{\mathcal{S}_t}]_{-ll}/[\widehat{\Theta}^{\mathcal{S}_t}]_{ll}$ for any $l \in \widehat{\mathcal{S}}_t$, where $\mathbf{X}_{\widehat{\mathcal{S}}_t\backslash\{l\}}$ denotes the sample matrix corresponding to $\widehat{\mathcal{S}}_t\backslash\{l\}$. By Corollary 1, $\mathcal{A}_t$ can be estimated as

$$\widehat{\mathcal{A}}_t = \Big\{ l : \widehat{\mathbf{e}}_{l,\widehat{\mathcal{S}}_t} \text{ is tested to be independent with } \ x_k \text{ for any } k \in \widehat{\mathcal{S}}_t\backslash\{l\} \Big\}.$$

Let $\widehat{\mathcal{S}}_{t+1} = \mathcal{N}\backslash\{\cup_{d=0}^{t}\widehat{\mathcal{A}}_d\}$, then Corollary 2 implies that for each $l \in \widehat{\mathcal{A}}_t$,

$$\widehat{\text{pa}}_l = \Big\{ k \in \widehat{\mathcal{S}}_{t+1} : [\widehat{\Theta}^{\widehat{\mathcal{S}}_t}]_{lk} \neq 0 \Big\},$$

and $\hat{\beta}_{lk} = -[\widehat{\Theta}^{\widehat{\mathcal{S}}_t}]_{lk}/[\widehat{\Theta}^{\widehat{\mathcal{S}}_t}]_{ll}$, for any $k \in \widehat{\text{pa}}_l$. The procedure is repeated until $|\widehat{\mathcal{S}}_t| \leq 1$, and we set $\widehat{T} = t+1$ and $\widehat{\mathcal{A}}_t = \widehat{\mathcal{S}}_t$ if $|\widehat{\mathcal{S}}_t| = 1$, and $\widehat{T} = t$ otherwise.

The details of the developed DAG learning method is summarized in Algorithm 1.

Note that the performance of the proposed method relies on the accuracy of the precision matrix estimation in Step 2a and the independence test procedure in Step 2b. Many existing methods in

Algorithm 1

1: **Input:** $\mathbf{X} \in \mathcal{R}^{n\times p}$, $\widehat{\mathcal{S}} = \{1, ..., p\}$, $\widehat{\mathbf{B}} = \{\hat{\beta}_{ij}\}_{i,j=1}^{p} = \mathbf{0}_{p\times p}$ and $t = 0$.

2: **Repeat:** until $|\widehat{\mathcal{S}}| \leq 1$:

a. Estimate $\widehat{\mathbf{\Theta}}^{\widehat{\mathcal{S}}}$ and compute $\widehat{\mathbf{e}}_{l,\widehat{\mathcal{S}}}$ for any $l \in \widehat{\mathcal{S}}$;

b. Estimate $\widehat{\mathcal{A}}_t = \left\{l : \widehat{\mathbf{e}}_{l,\widehat{\mathcal{S}}} \perp\!\!\!\perp x_k, \text{ for any } k \in \widehat{\mathcal{S}}\backslash\{l\}\right\}$;

c. Estimate $\widehat{\text{pa}}_l = \{k \in \widehat{\mathcal{S}}\backslash\widehat{\mathcal{A}}_t : [\widehat{\Theta}^{\widehat{\mathcal{S}}}]_{kl} \neq 0\}$ and $\widehat{\beta}_{lk} = -[\widehat{\Theta}^{\widehat{\mathcal{S}}}]_{lk}/[\widehat{\Theta}^{\widehat{\mathcal{S}}}]_{ll}$ for any $l \in \widehat{\mathcal{A}}_t$ and $k \in \widehat{\text{pa}}_l$;

d. Let $\widehat{\mathcal{S}} = \widehat{\mathcal{S}}\backslash\widehat{\mathcal{A}}_t$ and $t \leftarrow t + 1$.

3: If $|\widehat{\mathcal{S}}| = 1$, set $\widehat{T} = t + 1$ and $\widehat{\mathcal{A}}_t = \widehat{\mathcal{S}}$; otherwise set $\widehat{T} = t$.

4: **Return:** $\{\widehat{\mathcal{A}}_t\}_{t=0}^{\widehat{T}-1}$ and $\widehat{\mathbf{B}}$.

literature can be adopted, such as the graphical Lasso algorithm (Friedman et al., 2008; Ravikumar et al., 2011) or the constrained sparse estimation method (Cai et al., 2011) for precision matrix estimation, and the Hilbert-Schmidt independence criterion (Gretton et al., 2008), the distance covariance measure (Székely et al., 2007; Székely and Rizzo, 2009) or the ball divergence (Pan et al., 2018) for independence test. For illustration, we adopt the graphical Lasso algorithm and the distance covariance measure in the proposed method, which yields satisfactory performance in all the numerical experiments in Section 7.

It is also worthy pointing out that the conditional independence test error is inevitable in practice and may affect the numerical performance of the proposed method. Once Type I error occurs, some nodes belonging to the current layer may be assigned to some upper layer; once Type II error occurs, some nodes belonging to some upper layer may be assigned to the current layer. Both errors will lead to reconstruction errors and reduce the DAG recovery accuracy. Moreover, if no node is assigned to $\mathcal{A}_0$ owing to Type I error, Algorithm 1 will return a null graph. To prevent it from happening, we may slightly modify Algorithm 1 by adaptively decreasing the significance level of the conditional independence test.

5. Statistical guarantees

In this section, we establish asymptotic consistency of the proposed method with the graphical Lasso and distance covariance measure in terms of exact DAG recovery. The consistency results are established with explicit dependence on the sample size n, the number of nodes p, the maximum cardinality of the Markov blankets $d = \max_{l\in\mathcal{N}} |\text{mb}_l|$ and the number of layers T. Note that the support of $\mathbf{\Theta}_{-ll}$ is a subset of the Markov blanket of node l (Park et al., 2021).

For simplicity, denote $\sigma^2_{max} = \max_{l\in\mathcal{N}} \sigma^2_l$, $\sigma^2_{min} = \min_{l\in\mathcal{N}} \sigma^2_l$, $\beta_{max} = \max_{l\in\mathcal{N},k\in\text{pa}_l} |\beta_{lk}|$ and $\beta_{min} = \min_{l\in\mathcal{N},k\in\text{pa}_l} |\beta_{lk}|$. Further, denote $f(n) = \Omega(g(n))$ if there exists a positive constant a such that $f(n) \geq ag(n)$ for all sufficiently large n. The following technical assumptions are made to establish the exact DAG recovery.

Assumption 1 *There exists some constant $\psi \in (0, 1]$ such that*

$$\max_{t\in\{0,...,T-1\}} \max_{r\in\mathcal{C}_t^c} \|\Gamma_{r\mathcal{C}_t}(\Gamma_{\mathcal{C}_t\mathcal{C}_t})^{-1}\|_1 \leq 1 - \psi,$$

where $\Gamma = \mathbf{\Sigma} \otimes \mathbf{\Sigma}$ *with* $\otimes$ *denoting the Kronecker product,* $\Gamma_{(l,k),(j,m)} := \Gamma_{lp+k,jp+m} = \Sigma_{lk}\Sigma_{jm}$, $\mathcal{C}_t = \{(l,k) \in \mathcal{S}_t \times \mathcal{S}_t : [\Theta^{\mathcal{S}_t}]_{lk} \neq 0\}$ *and* $\mathcal{C}_t^c = (\mathcal{S}_t \times \mathcal{S}_t)\backslash\mathcal{C}_t$.

Assumption 2 *For any* $j \in \mathcal{N}$, ϵ_j/σ_j *follows a sub-Gaussian distribution with parameter* γ.

Assumption 1 limits the correlation between the zero and non-zero elements in Γ, which is analogous to the irrepresentable condition in Zhang and Yu (2006) or the incoherence condition in Ravikumar et al. (2011). Assumption 2 characterizes the noise distribution and implies that $x_j/\sqrt{\Sigma_{jj}}$ also follows a sub-Gaussian distribution with parameter γ.

Lemma 2 *Suppose Assumptions 1 and 2 hold, and* $n = \Omega(d^2 \log p)$. *For any* $l \in \mathcal{N}$, *there exist some positive constants* a_1, a_2 *and* $\tau > 4$ *such that with probability at least* $1 - a_2 p^{2-\tau}$, *there holds*

$$\left\| \frac{\widehat{\mathbf{\Theta}}_{\cdot l}}{\widehat{\Theta}_{ll}} - \frac{\mathbf{\Theta}_{\cdot l}}{\Theta_{ll}} \right\|_2 \leq a_1 \gamma^2 \tau^{1/2} \sqrt{\frac{d \log p}{L_2^2 n}}, \tag{4}$$

provided that the regularization parameter in estimating $\mathbf{\Theta}$ *is* $\lambda_{n,0} \propto \sqrt{\log p/n}$, *where* $L_2 = \sigma_{max}^{-2} \min\left\{1, (\frac{\sigma_{min}}{\sigma_{max}})^2 L_1^{-1}\right\}$ *with* $L_1 = \beta_{max}(1 + d\beta_{max})$.

Lemma 2 paves a bridge between the estimated precision matrix and the estimated residuals, and plays a crucial role in establishing the consistency for the exact DAG recovery.

Assumption 3 *There exists a positive constant* λ_{max} *such that* $\Lambda_{max}\big(\frac{1}{n}\mathbf{X}^T\mathbf{X}\big) \leq \lambda_{max}$, *where* $\Lambda_{max}(\cdot)$ *denotes the maximum eigenvalue of a matrix.*

Assumption 4 *For any* $t = 0, ..., T-1$, *we have*

$$\max_{k \in \mathcal{S}_t \backslash \{l\}} dcov^2(e_{l,\mathcal{S}_t}, x_k) \begin{cases} = 0, & \text{if } l \in \mathcal{A}_t; \\ \geq \rho_{n,t}^2, & \text{if } l \in \mathcal{S}_t \backslash \mathcal{A}_t, \end{cases}$$

where $\rho_{n,t}^2 = \Omega(\max\{d^3 n^{-1/2} \log^{1/2}(\max\{|\mathcal{S}_t|, n\}), n^{-\eta}\})$ *with* $0 < \eta < \frac{1}{2}$, *and* $|\mathcal{S}_t|$ *denotes the cardinality of* $\mathcal{S}_t$.

Assumption 3 is a standard regularity condition on the sample covariance matrix of $\mathbf{X}$ in $\mathcal{N}$, which also regulates the sample covariance matrix of $\mathbf{X}_{\mathcal{S}_t}$ since $\Lambda_{max}\big(\frac{1}{n}\mathbf{X}_{\mathcal{S}_t}^T\mathbf{X}_{\mathcal{S}_t}\big) \leq \Lambda_{max}\big(\frac{1}{n}\mathbf{X}^T\mathbf{X}\big)$ for any $t = 0, ..., T-1$. Assumption 4 assures that the distance covariance is sufficient in discriminating nodes in $\mathcal{A}_t$ or $\mathcal{S}_t\backslash\mathcal{A}_t$. Similar assumptions have also been employed in Kalisch and Bühlmann (2007) and Ha et al. (2016).

Theorem 2 (Consistency of $\widehat{\mathcal{A}}_0$) *Suppose that all the assumptions in Lemma 2 and Assumptions 3 and 4 hold, and* $n = \Omega(d^6 \log p)$. *Then there exist some positive constants* a_3, a_4, a_5 *and* a_6 *such that*

$$P(\widehat{\mathcal{A}}_0 = \mathcal{A}_0) \geq 1 - a_3 p^{4-\tau} - a_4 p^2 \exp\{-a_5 n^{(1-2\eta)/3}\},$$

provided that the significance level of the independence test is set as $\alpha_n = 2(1 - \Phi(a_6\sqrt{n}\rho_{n,0}))$, *where* $\Phi(\cdot)$ *denotes the distribution function of* $N(0,1)$.

Theorem 2 shows that the lowest layer $\mathcal{A}_0$ in $\mathcal{G}$ can be exactly recovered by the proposed method with properly chosen significance level α_n. Note that the chosen α_n depends on the lower bound of distance covariance defined in Assumption 4, and it is clear that $\alpha_n \to 0$ as $n \to \infty$. Similar choice of α_n has also been employed in Kalisch and Bühlmann (2007). After reconstructing $\mathcal{A}_0$, the selection consistency of the parent set for nodes in $\mathcal{A}_0$ can also be established, following from the fact that $\min_{l\in\mathcal{A}_0,k\in\text{pa}_l}|\Theta_{lk}| = \min_{l\in\mathcal{A}_0,k\in\text{pa}_l}\sigma_l^{-2}|\beta_{lk}| \geq \sigma_{max}^{-2}\beta_{min}$ and a similar treatment in Theorem 2 of Ravikumar et al. (2011).

Corollary 3 *Suppose that all the assumptions in Theorem 2 are satisfied, and* $n = \Omega((d^6 + \beta_{min}^{-2})\log p)$. *Then there exists some positive constant* a_7 *such that*

$$P\Big(\big\{\widehat{pa}_l = pa_l : l \in \widehat{\mathcal{A}}_0\big\}\Big|\widehat{\mathcal{A}}_0 = \mathcal{A}_0\Big) \geq 1 - a_7 p^{2-\tau}.$$

Further, let $\widehat{\mathcal{S}}_1 = \mathcal{N}\backslash\widehat{\mathcal{A}}_0$, and we apply the similar treatment on $\widehat{\mathcal{S}}_1$ to establish consistency in estimating $\widehat{\mathcal{A}}_1$, as well as other upper layers. In the spirit of mathematical induction, we arrive at the following theorem on the asymptotic estimation consistency of $\widehat{\mathcal{G}}$.

Theorem 3 (Consistency of $\widehat{\mathcal{G}}$) *Suppose that all the assumptions in Corollary 3 are satisfied, and* $n = \Omega\big(T^{1/(\tau-4)}(d^6 + \beta_{min}^{-2})(\log(\max\{p,n\}))^{3/(1-2\eta)}\big)$. *Then there holds*

$$P(\widehat{\mathcal{G}} = \mathcal{G}) \longrightarrow 1, \text{ as } n \to \infty,$$

provided that the regularization parameter in estimating $\mathbf{\Theta}^{\mathcal{S}_t}$ *is* $\lambda_{n,t} \propto \sqrt{\log(\max\{|\mathcal{S}_t|, n\})/n}$.

Theorem 3 ensures that the linear non-Gaussian DAG $\mathcal{G}$ can be consistently recovered by the proposed method even under the high dimensional setting. Theorem 3 can be further extended by relaxing the sub-Gaussian noise assumption, such as a noise distribution with $(4m)$-th bounded moment; that is, $\max_j E((\epsilon_j/\sigma_j)^{4m}) \leq K_m$ for an integer $m > 1$ and a positive constant K_m. Note that this bounded moment assumption further implies that $x_j/\sqrt{\Sigma_{jj}}$ also has $(4m)$-th bounded moment (Ghoshal and Honorio, 2018).

Corollary 4 *Suppose that all the assumptions in Theorem 3 are satisfied, except that Assumption 2 is relaxed to a noise distribution with* $(4m)$*-th bounded moment, Assumption 4 holds with* $\rho_{n,t}^2 = \Omega\big(\max\{d^3 n^{-\frac{1}{2}}|\mathcal{S}_t|^{\frac{2}{m}}(\max\{|\mathcal{S}_t|, n\})^{\frac{\tau-4}{2m}}, n^{-\eta}\}\big)$ *for some constants* $4 < \tau < m + 4$ *and* $0 < \eta < \frac{1}{2}$, *and* $n = \Omega\big(T^{\frac{1}{\min\{\tau-4, 2m\phi-1\}}}(d^6 + \beta_{min}^{-2})p^{\max\{\frac{4}{m}, \frac{2}{2m\phi-1}\}}(\max\{p,n\})^{\frac{\tau-4}{m}}\big)$ *for some constant* $\frac{1}{2m} < \phi < \frac{1}{2} - \eta$ *and* $m > 1$. *Then, there holds*

$$P(\widehat{\mathcal{G}} = \mathcal{G}) \longrightarrow 1, \text{ as } n \to \infty,$$

provided that $\lambda_{n,t} \propto |\mathcal{S}_t|^{\frac{2}{m}} n^{-\frac{m+4-\tau}{2m}}$.

Corollary 4 establishes the asymptotic DAG recovery of the proposed method with $(4m)$-th bounded moment noise distributions, which requires a relatively larger sample size compared with that in Theorem 3. Also, it is worthy pointing out that the sample complexities in Theorem 3 and Corollary 4 are both polynomial in d and T. Thus, under high-dimensional settings with $n \ll p$, d and T should be relatively small indicating that the DAGs need to be relatively sparse and shallow. More importantly, both Theorem 3 and Corollary 4 are established without assuming the parental faithfulness assumption in Wang and Drton (2020) or the ordered noise variance assumption in Ghoshal and Honorio (2018), indicating a more general applicability of the proposed method.

6. Related work

In this section, we compare the proposed method, denoted as TL, with some existing competitors in terms of their computational complexity and theoretical results. Specifically, we compare TL with the direct high-dimension learning algorithm (MDirect; Wang and Drton, 2020), the pairwise learning algorithm (Pairwise; Hyvarinen and Smith, 2013), the direct learning algorithm (Direct; Shimizu et al., 2011), the ICA-based learning algorithm (ICA; Shimizu et al., 2006). Note that all the afore-mentioned methods estimate the causal orderings in a top-down fashion, whereas TL reconstructs DAG in a bottom-up fashion by leveraging of the concept of topological layer. Most recently, following a different line of research, some works on linear structural equation models (Ghoshal and Honorio, 2018; Park, 2020; Park et al., 2021) can also be employed to learn linear non-Gaussian DAG under some additional ordered noise variance assumption. In particular, we also compare TL with the bottom-up search LISTEN algorithm (Ghoshal and Honorio, 2018).

The computational complexity of TL is largely determined by the precision matrix estimation and the independence tests in Steps 2a and 2b of Algorithm 1. The complexity of Step 2a using graphical Lasso (Friedman et al., 2008) is of order $O\big(n|\mathcal{S}_t|^2+R|\mathcal{S}_t|^3\big)$ for both low and high dimensional settings, where $|\mathcal{S}_t|$ denotes the cardinality of $\mathcal{S}_t = \sum_{k=t}^{T-1} \mathcal{A}_k$ and R denotes the number of coordinate descent cycles until convergence. The computational complexity of Step 2b using the distance covariance measure (Székely et al., 2007; Huo and Székely, 2016) is of order $O(n \log n|\mathcal{S}_t|^2)$. Therefore, the computational complexity of Algorithm 1 in learning a random linear non-Gaussian DAG with T layers is of order $O\big(\sum_{t=0}^{T-1}\big(R|\mathcal{S}_t|^3 + n\log n|\mathcal{S}_t|^2\big)\big)$. Theoretically, TL does not require the popularly-adopted faithfulness assumption (Uhler et al., 2013; Peters et al., 2017), the parental faithfulness assumption (Wang and Drton, 2020) or the ordered noise variance assumption (Ghoshal and Honorio, 2018) as shown in Corollaries 1 and 2, and the DAG recovery consistency of TL, established in Theorem 3 and Corollary 4, guarantees that the linear non-Gaussian DAG can be exactly recovered by TL with high probability under the high-dimensional setting.

MDirect is one of the most recently proposed learning methods in literature, whose computational complexity in the worst-case scenario is at least of order $O(J^2(n+J)p^{J+1})$, where J denotes the maximum in-degree of the DAG. It is worthy pointing out that MDirect is mainly designed for high-dimensional setting and it requires to search all the possible subset of size no more than J for computing the determining statistic sequentially. Thus, J is suggested to be small in Wang and Drton (2020). Theoretically, MDirect requires the parental faithfulness assumption, which is weaker than the faithfulness assumption, but is still somewhat restrictive as illustrated in Section 3.2 and is usually difficult to verify in practice. Under the parental faithfulness condition as well as other regularity conditions, the DAG recovery consistency is also established for MDirect in Wang and Drton (2020).

The computational complexity of ICA is dominated by Steps 1 and 5 of Algorithm A in Shimizu et al. (2006), and thus its order is up to $O(np^3 + p^4)$. The computational complexity of Pairwise is also of order $O(np^3 + p^4)$. Moreover, the dominant computational part of Direct is the employed kernel-based independence measure, and thus its total computational complexity is of order $O(np^3M^2 + p^4M^3)$, where M denotes the maximal rank obtained by the applied low-rank matrix decomposition in the kernel-based independence measure. Note that these three methods are mostly algorithm based, and their theoretical properties remain undeveloped. These methods may also suffer from heavy computational cost even when dealing with a medium-sized DAG. For example, as

reported in Section 4 of Shimizu et al. (2011), the median computational time of Direct for a sparse DAG with 500 observations and 100 nodes is up to 2.35 hours.

LISTEN recovers the causal ordering among the nodes in a bottom-up fashion. Particularly, given the last identified node, LISTEN updates the corresponding elements in the estimated precision matrix and determines the next node by checking the smallest diagonal values of the updated precision matrix. Its computational complexity is of order $O(Lp^5 + np^2)$ for a dense DAG, where np^2 comes from computing the sample covariance matrix, Lp^5 comes from the precision matrix estimation by using CLIME (Cai et al., 2011), and L denotes the number of arithmetic operations performed (Spielman and Teng, 2003) in CLIME. When the DAG is sparse, the computational complexity of LISTEN can reduce to $O(L(p^4 + pd^4) + np^2)$, where $O(Lp^4)$ comes from solving p linear programmings in CLIME each with complexity up to $O(Lp^3)$ (Spielman and Teng, 2003), and $O(Lpd^4)$ comes from updating only the elements corresponding to the Markov blanket of the last identified node in each iteration. Theoretically, under the ordered noise variance assumption as well as other regularity conditions, the exact DAG recovery can also be established for LISTEN (Ghoshal and Honorio, 2018).

For easy reference, we summarize the computational complexities and theoretical results mentioned above in the following table.

	Complexity	Faithfulness?	Consistency
TL	$O\big(\sum_{t=0}^{T-1}\big(R\lvert\mathcal{S}_t\rvert^3 + n\log n\lvert\mathcal{S}_t\rvert^2\big)\big)$	No	✓
MDirect	$O(J^2(n+J)p^{J+1})$	Parental faithfulness	✓
ICA	$O(np^3 + p^4)$	No	−
Direct	$O(np^3M^2 + p^4M^3)$	No	−
Pairwise	$O(np^3 + p^4)$	No	−
LISTEN	$O(L(p^4 + pd^4) + np^2)$	No	✓

Table 1: A comparison summary of all competing methods. Here "−" denotes that theoretical results including DAG recovery consistency are unavailable.

From Table 1, it is clear that TL is computationally more efficient than the other competing methods, especially for the sparse or shallow graphs with a large number of nodes. When the DAG is relatively sparse, R can be treated as a constant, and the computational complexity of TL becomes $O(p^4 + np^3\log n)$ in the worst-case scenario with $T = p$. When the DAG is a shallow hub graph with $T = 2$, the computational complexity of TL becomes $O(p^3 + np^2\log n)$, where the $\log n$ term may be discarded if an alternative method of independence test (Jitkrittum et al., 2017) is employed. It is also interesting to notice that the computational complexity of MDirect is of an exponential order in J, and thus it may suffer from serious computational challenges when some nodes have a relatively large number of parents. The computational complexity of LISTEN is comparable to that of TL for the dense graphs. The computational advantage of TL is also supported by the numerical results in Section 7. For instance, TL is the only method that can produce a reasonable estimate of a shallow DAG graph with 1000 nodes. Detailed numerical comparisons in term of computational complexity are also provided in Section 7.1.

Furthermore, TL and MDirect are two comparable methods in the literature of learning linear non-Gaussian DAG, with asymptotic DAG recovery consistency based on the property of non-Gaussianality only. Particularly, Theorem 3 implies that $\log p$ in TL is allowed to diverge at order $o(n^{(1-2\eta)/3})$ with $0 < \eta < 1/2$, whereas Corollary 3 in Wang and Drton (2020) shows that $\log p$

in MDirect can only diverge at order $o(n^{1/2K})$ with $K \geq 3$. It is clear that the order of $\log p$ in TL can be much larger than that in MDirect, and more importantly, we want to emphasize again that the parental faithfulness assumption in Wang and Drton (2020) is unnecessary for Theorem 3 and Corollary 4. On the other hand, the asymptotic DAG recovery can also be established for LISTEN, but based on a different set of assumptions on the ordered noise variances, which can be difficult to verify in practice.

7. Numerical experiments

In this section, we examine the numerical performance of TL, and compare it against some popular learning linear non-Gaussian DAG methods, including the afore-mentioned MDirect, Pairwise and ICA, and some general DAG learning methods, including LISTEN, the high dimensional constraint-based PC algorithm (PC; Kalisch and Bühlmann, 2007) and a hybrid version of max-min hill climbing algorithm (MMHC; Tsamardinos et al., 2006). Particularly, TL adopts the graphical Lasso algorithm with a fixed regularization parameter as suggested in Section 5, and the independence test based on distance covariance measure. MDirect is implemented in the `R` package `highDLingam` (Wang and Drton, 2020), both ICA and MMHC are implemented in the `R` package `CompareCausalNetworks`, and LISTEN is implemented by slightly modifying the `R` package `EqVarDAG`. Furthermore, we implement Pairwise by using the `R` package `causalXtreme` (Gnecco et al., 2021), and further extend it with the Lasso algorithm for DAG in high dimensional cases, and we implement PC by using the `R` package `pcalg` (Kalisch et al., 2012), which outputs a partial DAG, and then we apply the treatment in Yuan et al. (2019) to convert it to a DAG by using the `pdag2dag` routine in the R package `pcalg`. Following the same treatment as in Kalisch and Bühlmann (2007), the significance level of independent tests in both TL and PC is set as $\alpha_n = 0.01$, and the least square estimation is also applied to estimate the connection strength of directed structures for MDirect, MMHC and PC.

The numerical performance of all the methods is evaluated in terms of estimation accuracy of directed edges and coefficients. For the accuracy of estimated directed edges, we employ the true positive rate (TPR) and false discovery rate (FDR) as the evaluation metric. To evaluate the closeness of the estimated and true DAG, we report the normalized structural Hamming distance (Tsamardinos et al., 2006), which measures the smallest number of edge insertions, deletions, and flips to convert the estimated DAG into the truth. For overall accuracy of the estimated DAG structure, we use the Matthews correlation coefficient (MCC) as an overall evaluation metric, which is also considered in Yuan et al. (2019). For evaluation of the coefficient estimation, we report the relative error between the estimated adjacency matrix $\widehat{\mathbf{B}}$ and the true adjacency matrix $\mathbf{B}$ in Frobenius norm that rel-Fnorm $:= \|\widehat{\mathbf{B}} - \mathbf{B}\|_F / \|\mathbf{B}\|_F$. Note that a good estimation is implied with small values of FDR, HM and rel-Fnorm, but large values of TPR and MCC.

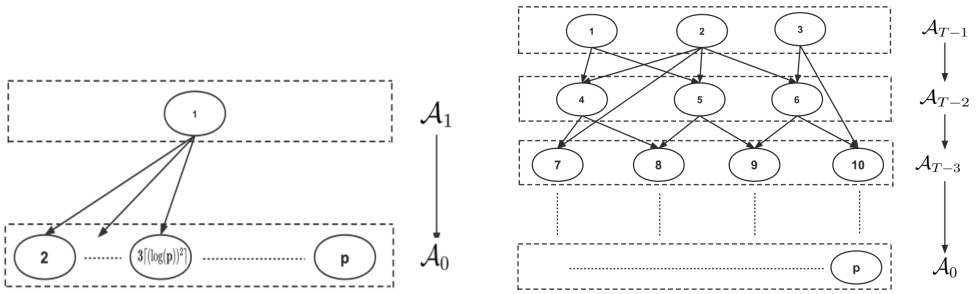
7.1 Simulated examples

In this section, the numerical performance of all the methods are evaluated in two simulated examples, where Example 1 considers a sparse hub graph, and Example 2 considers a scale-free graph generated by the Barabási-Albert (BA) model.

Example 1. We consider a sparse hub graph with $T = 2$, $\mathcal{A}_0 = \{2, ..., p\}$ and $\mathcal{A}_1 = \{1\}$, whose DAG structure is illustrated in Figure 2a. Specifically, we select the first $3\lceil(\log p)^2\rceil$ nodes in $\mathcal{A}_0$ as children of node 1 and the remaining nodes are isolated. Moreover, we randomly draw

the noise terms from various distributions, including uniform distribution on $[-3, 3]$, student t distribution with 9 degree of freedom, and double exponential distribution with location parameter 0 and scale parameter $\sqrt{1.5}$, and the corresponding variance σ^2 of these noise terms are $3, 9/7$ and 3, respectively. Moreover, the coefficient of each directed edge is uniformly generated from $[-1.5, -0.5] \cup [0.5, 1.5]$.

Example 2. We consider a similar data generating scheme as in Wang and Drton (2020). Specifically, we start with a graph with only one node, and at each step, a node with 2 directed edges are added to the graph. The probability of generating a directed edge from each previous node to the newly added node is proportional to the number of neighbors of the previous node. Its DAG structure is illustrated in Figure 2b. Moreover, we generate each noise terms ϵ_k from $\upsilon_k \times \text{Uniform}[-3, 3]$ where $\upsilon_k \sim \text{Uniform}[1, 2]$ denotes a prior parameter and is fixed for each ϵ_k, and then the corresponding variance σ_k^2 of noise term ϵ_k is $3\upsilon_k^2$. Moreover, the coefficient of each directed edge is uniformly generated from $[-1.5, -0.5] \cup [0.5, 1.5]$.



(a) (b)

Figure 2: The topological layer of the DAG structures in Examples 1 and 2.

For each example, we repeat the data generating scheme 50 times and the averaged performance of all the methods under the cases with $(n, p) = (200, 100), (200, 200), (400, 200)$ and $(400, 1000)$ are summarized in Tables 2 and 3. Note that ICA is only designed for the low dimensional case with $p < n$, and some methods do not produce any results for cases with large p in Examples 1 and 2 after more than 48 hours.

It is evident from Tables 2 and 3 that TL outperforms all the other competitors in almost all the cases, except that it yields the second or third best TPR in the cases with $(n, p) = (200, 100)$ and $(400, 200)$ in Example 2. In these two cases, Pairwise and ICA attain higher TPR, largely due to the fact that they tend to produce very dense graphs with many false edges and thus have much higher FDR. Note that MDirect does not perform well, possible due to its sensitivity to the data generating scheme, and the performance of LISTEM appears less satisfactory largely due to the violation of its required ordered noise variance assumption. It is also interesting to point out that the performance of TL may be further improved with a finer tuning scheme, at the cost of increasing computational cost.

To scrutinize the exact recovery rate of the proposed method, we consider the same setting as Example 1 with noises generated from a mixed distribution, and a modified setting with noise generated solely from a uniform distribution on $[-3, 3]$. Note that all the noise terms in Example 1 have $(4m)$-th bounded moments. For example, the noise term ϵ_k following student t distribution with 9 degree of freedom has bounded 8-th moment that $E((\epsilon_k/\sigma_k)^8) = 7^4 < \infty$. Moreover, the

(n, p)	Method	TPR	FDR	MCC	HM	rel-Fnorm
$(200, 100)$	TL	0.9923 (0.0018)	0.0015 (0.0007)	0.9953 (0.0010)	0.0001 (0.0000)	0.0861(0.0045)
	MDirect	0.0498 (0.0018)	0.0132 (0.0002)	0.0032 (0.0006)	0.0475 (0.0005)	1.1212 (0.0024)
	Pairwise	0.8086 (0.0194)	0.8078 (0.0046)	0.3871 (0.0095)	0.0235 (0.0003)	0.8378 (0.0085)
	ICA	0.5569 (0.0049)	0.6977 (0.0020)	0.4048 (0.0029)	0.0114 (0.0001)	0.9265 (0.0013)
	MMHC	0.2277 (0.0010)	0.8148 (0.0013)	0.1994 (0.0011)	0.0117 (0.0001)	0.9214 (0.0016)
	PC	0.1631 (0.0012)	0.5969 (0.0049)	0.2529 (0.0023)	0.0071 (0.0000)	0.8947 (0.0013)
	LISTEN	0.2791 (0.0055)	0.1738 (0.0123)	0.4774 (0.0071)	0.0051 (0.0001)	0.8337 (0.0046)
$(200, 200)$	TL	0.9805 (0.0050)	0.0002 (0.0002)	0.9899 (0.0026)	0.0000 (0.0000)	0.0971 (0.0070)
	MDirect	0.0277 (0.0011)	0.9972 (0.0001)	0.0021 (0.0003)	0.0232 (0.0002)	1.1287 (0.0028)
	Pairwise	0.4559 (0.0253)	0.9348 (0.0042)	0.1666 (0.0104)	0.0195 (0.0005)	0.7793 (0.0157)
	ICA	**	**	**	**	**
	MMHC	0.1767 (0.0009)	0.9355 (0.0004)	0.1036 (0.0006)	0.0073 (0.0000)	0.9935 (0.0018)
	PC	0.1212 (0.0011)	0.8516 (0.0019)	0.1322 (0.0014)	0.0034 (0.0000)	0.9429 (0.0013)
	LISTEN	0.0460 (0.0123)	0.9914 (0.0023)	0.0150 (0.0053)	0.0134 (0.0002)	1.0741 (0.0078)
$(400, 200)$	TL	0.9979 (0.0006)	0.0000 (0.0000)	0.9980 (0.0003)	0.0000 (0.0000)	0.0537 (0.0022)
	MDirect	0.0227 (0.0006)	0.9977 (0.0001)	0.0004 (0.0002)	0.0235 (0.0001)	1.1071 (0.0016)
	Pairwise	0.8388 (0.0140)	0.7512 (0.0041)	0.4547 (0.0075)	0.0057 (0.0000)	0.7505 (0.0068)
	ICA	0.7002 (0.0031)	0.8277 (0.0006)	0.3447 (0.0013)	0.0079 (0.0000)	0.9034 (0.0016)
	MMHC	**	**	**	**	**
	PC	**	**	**	**	**
	LISTEN	**	**	**	**	**
$(400, 1000)$	TL	0.9910 (0.0057)	0.0000 (0.0000)	0.9953 (0.0030)	0.0000 (0.0000)	0.0642 (0.0068)
	MDirect	**	**	**	**	**
	Pairwise	**	**	**	**	**
	ICA	**	**	**	**	**
	MMHC	**	**	**	**	**
	PC	**	**	**	**	**
	LISTEN	**	**	**	**	**

Table 2: The averaged measures of all the methods in Example 1 together with their standard errors in parentheses. Here ** denotes the fact that the corresponding methods are either not applicable or take too long to produce any results.

noise term drawn from $U[-3, 3]$ is also sub-Gaussian distributed. Following the similar analysis in Park et al. (2021), we replicate both cases with $p = 50$ or 100 for 50 times, and vary $n/p \in \{2, 4, ..., 12\}$ for the original setting and $n/\log(p) \in \{25, 50, ..., 200\}$ for the modified setting, respectively. The averaged exact recovery rates of the proposed method are displayed in Figure 3. It is clear that the exact recovery rates increase with ratio n/p or $n/\log(p)$, and eventually converge to 1 when the ratios are sufficiently large. This confirms the asymptotic consistencies of the proposed method in Theorem 3 and Corollary 4 for a sub-Gaussian noise or a $(4m)$-th bounded moment noise.

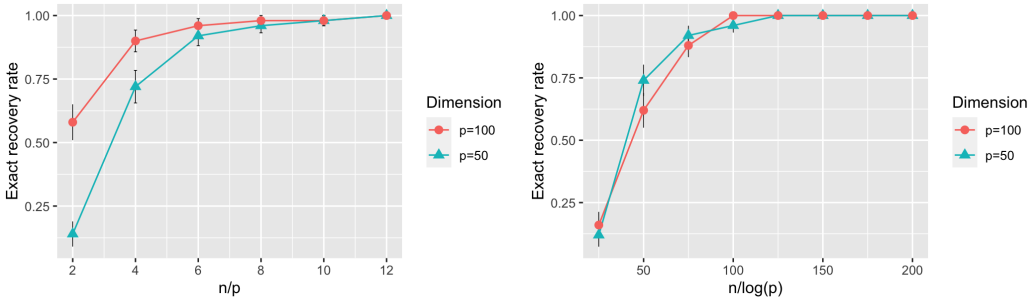
Furthermore, the averaged running times of all the competing methods in Examples 1 and 2 are reported in Table 4. It is clear that TL has remarkable computational efficiency compared with its competitors, especially when the underlying DAG has some shallow structure. In Example 1, the averaged running time of TL is much shorter than the other method, and it is the only method that can produce a reasonable estimate of a DAG with 1000 nodes. In Example 2, the averaged running time of TL is shorter than that of MDirect and Pairwise, and comparable to that of ICA, MMHC, PC and LISTEN.

7.2 Spread of COVID-19

We now apply TL to analyze the spread of COVID-19 based on the daily global confirmed cases collected by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University,

(n, p)	Method	TPR	FDR	MCC	HM	rel-Fnorm
$(200, 100)$	TL	0.7113 (0.0098)	0.1379 (0.0072)	0.7788 (0.0083)	0.0080 (0.0003)	0.7518 (0.0247)
	MDirect	0.1266 (0.0016)	0.9141 (0.0012)	0.0822 (0.0014)	0.0444 (0.0002)	1.1210 (0.0037)
	Pairwise	0.8401 (0.0102)	0.7931 (0.0026)	0.3969 (0.0052)	0.0674 (0.0005)	0.8839 (0.0077)
	ICA	0.7056 (0.0028)	0.7034 (0.0014)	0.4413 (0.0013)	0.0394 (0.0003)	0.8339 (0.0029)
	MMHC	0.3027 (0.0023)	0.3421 (0.0046)	0.4391 (0.0032)	0.0170 (0.0001)	0.8882 (0.0027)
	PC	0.2007 (0.0021)	0.5656 (0.0039)	0.2859 (0.0028)	0.0211 (0.0001)	0.9454 (0.0020)
	LISTEN	0.3779 (0.0071)	0.2330 (0.0061)	0.5315 (0.0064)	0.0147 (0.0002)	0.8401 (0.0058)
$(200, 200)$	TL	0.7000 (0.0092)	0.1554 (0.0043)	0.7663 (0.0066)	0.0043 (0.0001)	0.7804 (0.0303)
	MDirect	0.0931 (0.0013)	0.9389 (0.0007)	0.0639 (0.0009)	0.0233 (0.0001)	1.1331 (0.0033)
	Pairwise	0.6303 (0.0089)	0.8929 (0.0013)	0.2439 (0.0034)	0.0560 (0.0004)	1.2092 (0.0180)
	ICA	**	**	**	**	**
	MMHC	0.2991 (0.0023)	0.3253 (0.0042)	0.4457 (0.0030)	0.0084 (0.0000)	0.8823 (0.0028)
	PC	0.1893 (0.0023)	0.5855 (0.0044)	0.2753 (0.0032)	0.0107 (0.0000)	0.9551 (0.0020)
	LISTEN	0.3097 (0.0041)	0.3146 (0.0060)	0.4571 (0.0046)	0.0083 (0.0001)	0.8912 (0.0047)
$(400, 200)$	TL	0.7701 (0.0066)	0.1168(0.0035)	0.8227 (0.0044)	0.0033 (0.0001)	0.6969 (0.0194)
	MDirect	0.0935 (0.0008)	0.9366 (0.0006)	0.0656 (0.0006)	0.0229 (0.0001)	1.1270 (0.0015)
	Pairwise	0.9403 (0. 022)	0.7772 (0.0009)	0.4488 (0.0013)	0.0335 (0.0001)	0.7579 (0.0029)
	ICA	0.7827 (0.0014)	0.7788 (0.0006)	0.4063 (0.0006)	0.0298 (0.0001)	0.8158 (0.0009)
	MMHC	0.3229 (0.0016)	0.3364 (0.0034)	0.4592 (0.0022)	0.0084 (0.0000)	0.8752 (0.0019)
	PC	0.2005 (0.0021)	0.6142 (0.0040)	0.2729 (0.0029)	0.0112 (0.0001)	0.9559 (0.0022)
	LISTEN	0.5233 (0.0081)	0.1394 (0.0045)	0.6678 (0.0061)	0.0056 (0.0001)	0.7372 (0.0070)
$(400, 1000)$	TL	0.7411 (0.0097)	0.1532 (0.0043)	0.7908 (0.0058)	0.0008 (0.0000)	0.6898 (0.0174)
	MDirect	0.0500 (0.0003)	0.9652 (0.0002)	0.0394 (0.0002)	0.0047 (0.0000)	1.1410 (0.0012)
	Pairwise	**	**	**	**	**
	ICA	**	**	**	**	**
	MMHC	0.2973 (0.0011)	0.3173 (0.0017)	0.4498 (0.0014)	0.0017 (0.0000)	0.8874 (0.0015)
	PC	0.1575 (0.0008)	0.6838 (0.0015)	0.2220 (0.0011)	0.0024 (0.0000)	0.9794 (0.0013)
	LISTEN	**	**	**	**	**

Table 3: The averaged measures of all the methods in Example 2 together with their standard errors in parentheses. Here ** denotes the fact that the corresponding methods are either not applicable or take too long to produce any results.



(a) $(4m)$-th bounded moment noise (b) Sub-Gaussian noise

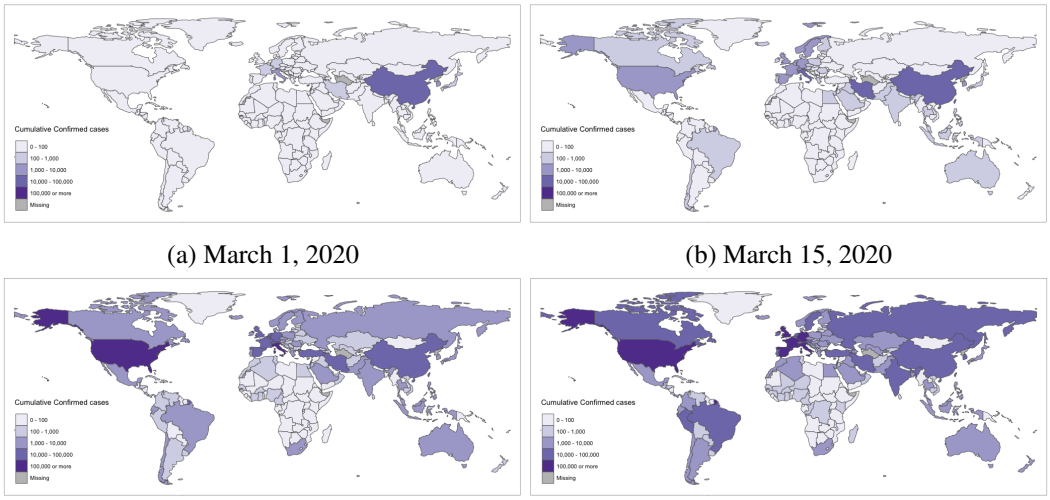
Figure 3: The averaged exact recovery rates of the proposed method over 50 replications and their corresponding standard errors.

which is publicly available at `https://github.com/CSSEGISandData/COVID-19`. Figure 4 displays the heat maps for the global cumulative confirmed cases for countries around the world from March 1st, 2020 to April 15th, 2020. It is clear that most confirmed cases of COVID-19 are first reported in China, Europe and Iran, then it quickly spreads to Middle East and North America, and finally most of the countries are affected by COVID-19, especially USA and west European

	(n, p)	TL	MDirect	Pairwise	ICA	MMHC	PC	LISTEN
Example 1	$(200, 100)$	0.18 (0.01)	0.38 (0.01)	0.24 (0.0)	0.09 (0.00)	0.33 (0.02)	1.66 (0.06)	1.52 (0.03)
	$(200, 200)$	0.26 (0.01)	4.06 (0.10)	2.00 (0.02)	**	0.59 (0.03)	4.42 (0.14)	156.77 (5.37)
	$(400, 200)$	0.77 (0.01)	8.42 (0.08)	4.56 (0.03)	1.37 (0.01)	**	**	**
	$(400, 1000)$	11.73 (0.41)	**	**	**	**	**	**
Example 2	$(200, 100)$	0.21 (0.01)	0.14 (0.00)	0.18 (0.00)	0.12 (0.00)	0.02 (0.00)	0.06 (0.00)	0.10 (0.01)
	$(200, 200)$	0.33 (0.02)	0.72 (0.02)	0.89 (0.01)	**	0.08 (0.00)	0.25 (0.00)	2.05 (0.10)
	$(400, 200)$	1.40 (0.03)	1.09 (0.02)	2.51 (0.01)	1.56 (0.02)	0.14 (0.00)	0.44 (0.01)	2.42 (0.07)
	$(400, 1000)$	28.63 (1.23)	114.85 (2.45)	**	**	7.78 (0.07)	33.63 (0.16)	**

Table 4: The averaged running times (in mins) of various methods in Examples 1 and 2 together with their standard errors in parentheses. Here ** indicates that the corresponding methods are either not applicable or take too long to produce any results.

countries. Interestingly, compared with other continents, countries in Africa appear to be much less affected.



(a) March 1, 2020 (b) March 15, 2020

(c) March 30, 2020 (d) April 15, 2020

Figure 4: Heat map of the cumulative confirmed cases for countries around the world from March 1st, 2020 to April 15th, 2020.

It is interesting to note that DAG is an efficient tool to describe the spread of COVID-19, where a directed edge indicates the virus is spread from one country to the other. Although virus-spread may not be necessarily acyclic, DAG provides insightful information on the future infection tendency of virus-spread among the countries. We pre-process the dataset and exclude those countries or regions with no confirmed cases for more than 10 days during March 1st, 2020 to April 15th, 2020. This leads to the daily confirmed cases in $p = 99$ countries or regions for a total of 61 days. Further, we convert the actual number of daily confirmed cases to the percentage over all countries, and use a 3-day moving average of the percentages as the observation for each day. We then apply TL to estimate the DAG for the spread of COVID-19, with 99 nodes and 736 directed edges.

Figure 5 shows the top 25 hub nodes with the number of their child nodes in the estimated DAG for the spread of COVID-19, which consists of mostly Eastern Asian and Western Asian countries, and most European countries. This concurs with the heat map in Figure 4b that these countries

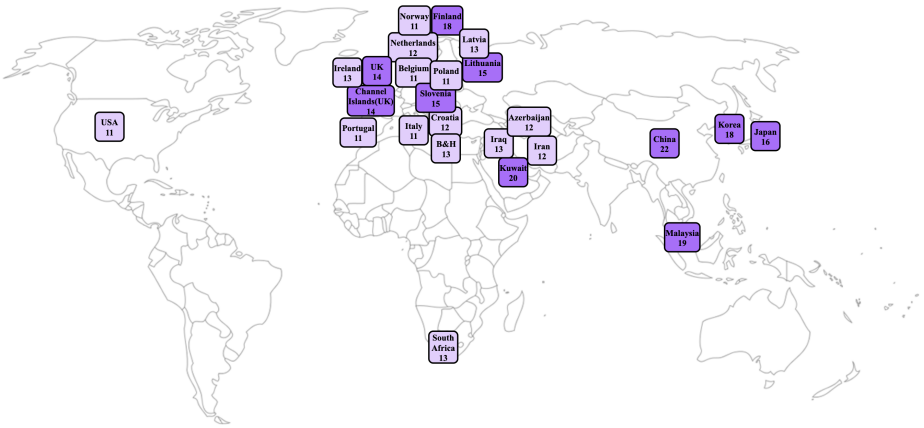


Figure 5: Top 25 hub nodes with the number of their child nodes in the estimated DAG for the spread of COVID-19.

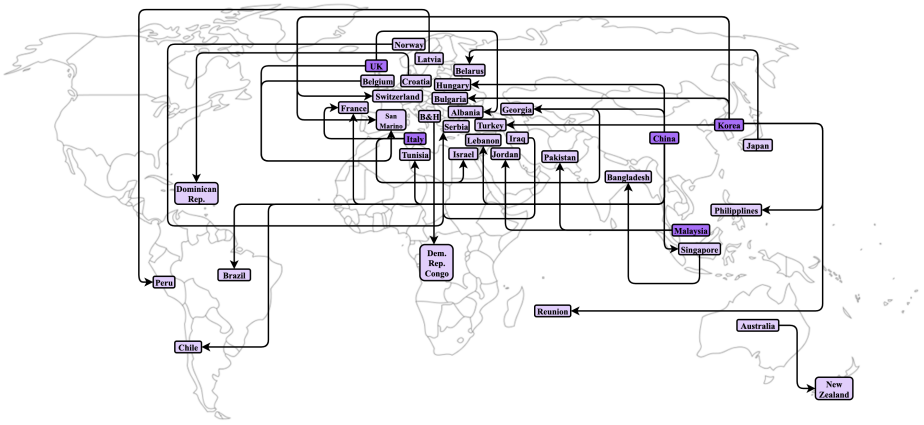


Figure 6: Top 30 directed edges in the estimated DAG for the spread of COVID-19.

reported many confirmed cases in middle March, and thus are more likely to spread the virus. Figure 6 presents 30 directed edges with largest estimated weights, showing that China, Korea, Italy and United Kingdom are the major countries that spread the virus to others. This trend of infection appear sensible, since these countries have more confirmed cases in early March as shown in Figure 4 and they are also closely connected to other countries due to their active economy or tourism attractions. Moreover, many directed edges are present among European countries, largely due to the fact that population movement and interaction in these countries are much more frequently than others. It is also interesting to note that Figure 6 shows no directed edges point from the United States of America and Canada to other countries, yet they are indeed hub nodes with 11 and 6 child nodes, respectively. Given the fact that there is a surge in the number of confirmed cases in these two countries as observed in Figure 4, one can expect the virus would spread from these two countries to their child nodes after April, 2020.

8. Discussion

This paper proposes an efficient method to learn linear non-Gaussian DAG in high dimensional cases with statistical guarantees. The proposed method leverages a concept of topological layers to facilitate DAG learning, which ensures that the parents of a node must belong to its upper layers, and thus naturally guarantees acyclicity. To learn the DAG, its layers can be reconstructed via precision matrix estimation and independence tests in a bottom-up fashion, and its parent-child relations can be directly obtained from the estimated precision matrix. More importantly, the proposed method can consistently recover the underlying DAG under more mild conditions than existing methods in literature. Its advantages over some popular competitors are also supported by numerical experiments on a variety of simulated and real-life examples.

Acknowledgments

The authors thank the Action Editor and the referee for their constructive suggestions, which significantly improve this paper. XH's research is supported in part by NSFC-11901375 and Shanghai Pujiang Program 2019PJC051, and JW's research is supported in part by GRF-11300919, GRF-11304520 and GRF-11206821.

Appendix A. Proof for Theorem 1

Proof of Theorem 1: First, if $l \in \mathcal{A}_0$, we have $e_{l,\mathcal{N}} = \epsilon_l$ by the SEM model in (1), and thus $e_{l,\mathcal{N}} \perp\!\!\!\perp x_k$ for any $k \in \mathcal{N}\backslash\{l\}$.

Next, we show that if $e_{l,\mathcal{N}} \perp\!\!\!\perp x_k$ for any $k \in \mathcal{N}\backslash\{l\}$, then $l \in \mathcal{A}_0$. If not, denote its youngest child as $j \in \mathrm{ch}_l$ such that node j belongs to the lowest layer compared with other nodes in ch_l, and it follows from (1) and (2) that

$$x_j = \sum_{k\in \mathrm{pa}_j} \beta_{jk} x_k + \epsilon_j = \sum_{k\in \mathrm{pa}_j} \beta_{jk}\big(\sum_{d\neq j} a_{kd}\epsilon_d\big) + \epsilon_j = \sum_{d\neq j}\big(\sum_{k\in \mathrm{pa}_j} \beta_{jk} a_{kd}\big)\epsilon_d + \epsilon_j, \tag{5}$$

where the second equality follows from the fact that $a_{kj} = 0$ for any $k \in \mathrm{pa}_j$.

The expected residual $e_{l,\mathcal{N}}$ can be obtained by

$$e_{l,\mathcal{N}} = x_l - \sum_{k\neq l} M_k^{(l)} x_k = x_l - \sum_{k\neq l, k\neq j} M_k^{(l)} x_k - M_j^{(l)} x_j, \tag{6}$$

where $M_k^{(l)}$ denotes the k-th element of $\mathbf{M}_{\mathcal{N}}^{(l)} = \mathbf{\Sigma}_{-l,-l}^{-1}\mathbf{\Sigma}_{-l,l} = -\mathbf{\Theta}_{-l,l}/\Theta_{l,l}$, and can be regarded as the partial correlation between nodes k and l given all other nodes in $\mathcal{N}$. Furthermore, we

decompose each term in (6) as

$$\begin{aligned} x_l &= \sum_{d\neq j} a_{ld}\epsilon_d, \\ M_j^{(l)} x_j &= \sum_{d\neq j} \Big(\sum_{k\in \text{pa}_j} M_j^{(l)} \beta_{lk} a_{kd} \Big)\epsilon_d + M_j^{(l)}\epsilon_j, \\ \sum_{k\neq l, k\neq j} M_k^{(l)} x_k &= \sum_{d\neq j} \Big(\sum_{k\neq l,k\neq j} M_k^{(l)} a_{kd} \Big)\epsilon_d + \Big(\sum_{k\neq l,k\neq j} M_k^{(l)} a_{kj} \Big)\epsilon_j \\ &= \sum_{d\neq j} \Big(\sum_{k\neq l,k\neq j} M_k^{(l)} a_{kd} \Big)\epsilon_d + \Big(\sum_{k\in \text{ de}_j} M_k^{(l)} a_{kj} \Big)\epsilon_j, \end{aligned}$$

where the first equality follows from (2) and the fact that $a_{lj} = 0$ since $j \in \text{de}_l$, the second equality follows from (5), and the last two equalities follow from (2), the fact that $l, j \notin \text{de}_j$ and $a_{kj} = 0$ for any $k \notin \text{de}_j$. Thus, (6) can be rewritten as

$$e_{l,\mathcal{N}} = \sum_{d\neq j} \Big(a_{ld} - \sum_{k\in \text{pa}_j} M_j^{(l)} \beta_{lk} a_{kd} - \sum_{k\neq l,k\neq j} M_k^{(l)} a_{kd} \Big)\epsilon_d - \Big(M_j^{(l)} + \sum_{k\in \text{ de}_j} M_k^{(l)} a_{kj} \Big)\epsilon_j. \quad (7)$$

where $M_j^{(l)}$ is nonzero due to the fact that node j is the youngest child of node l.

Now we consider two cases: (i) $\text{ch}_j = \emptyset$; and (ii) $\text{ch}_j \neq \emptyset$. For case (i), we have $\text{ch}_j = \emptyset$, which implies that $j \in \mathcal{A}_0$, and thus $\text{de}_j = \emptyset$. Then the coefficient of ϵ_j in (7) reduces to $M_j^{(l)}$. For case (ii), it implies that $j \in \bigcup_{t=1}^{T-1} \mathcal{A}_t$ and thus $\text{de}_j \neq \emptyset$. Note that for any $k \in \text{de}_j$, there holds that nodes l and k are independent given all the other nodes including node j, due to the fact that node j belongs to the lowest topological layer among all the nodes in ch_l. This immediately implies that $M_k^{(l)} = 0$ for any $k \in \text{de}_j$, and thus the coefficient of ϵ_j in (7) also reduces to $M_j^{(l)}$. Therefore, (7) reduces to

$$e_{l,\mathcal{N}} = \sum_{d\neq j} \Big(a_{ld} - \sum_{k\in \text{pa}_j} M_j^{(l)} \beta_{lk} a_{kd} - \sum_{k\neq l,k\neq j} M_k^{(l)} a_{kd} \Big)\epsilon_d - M_j^{(l)}\epsilon_j. \quad (8)$$

Putting (5) and (8) together, since $e_{l,\mathcal{N}} \perp\!\!\!\perp x_j$, it follows from Lemma 1 that ϵ_j must be Gaussian distributed, which contradicts with the generating scheme of the SEM model in (1). This completes the proof. ■

Appendix B. Proof for Lemma 2

Lemma 3 *Suppose that Assumptions 1 and 2 hold, and $n = \Omega(d^2 \log p)$, then there exist some positive constants b_1, b_2 and $\tau > 4$ such that with probability at least $1 - b_2 p^{2-\tau}$, there holds*

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{max} \leq b_1 \gamma^2 \tau^{1/2} \sqrt{\frac{\log p}{n}},$$

provided that $\lambda_{n,0} \propto \sqrt{\log p/n}$, where $\|\cdot\|_{max}$ denotes the matrix max norm.

Lemma 3 is a standard consistency result for precision matrix estimation by the graphical Lasso algorithm Ravikumar et al. (2011), and thus its proof is omitted here.

Proof of Lemma 2: Define

$$\mathcal{D}_1 = \Big\{ \Big\| \frac{\widehat{\mathbf{\Theta}}_{\cdot l}}{\widehat{\Theta}_{ll}} - \frac{\mathbf{\Theta}_{\cdot l}}{\Theta_{ll}} \Big\|_2 \geq 4b_1 \gamma^2 \tau^{1/2} \sqrt{\frac{d \log p}{L_2^2 n}} \Big\},$$

and $\epsilon_n = 2b_1 \sigma_{max}^2 \gamma^2 \tau^{1/2} \sqrt{\frac{\log p}{n}}$. Then, $P(\mathcal{D}_1)$ can be decomposed as

$$\begin{aligned} P(\mathcal{D}_1) &= P\Big(\mathcal{D}_1 \cap \big\{\sigma_{max}^2 \min_{l \in \mathcal{N}} |\widehat{\Theta}_{ll}| \leq 1 - \frac{1}{2}\epsilon_n\big\}\Big) + P\Big(\mathcal{D}_1 \cap \big\{\sigma_{max}^2 \min_{l \in \mathcal{N}} |\widehat{\Theta}_{ll}| \geq 1 - \frac{1}{2}\epsilon_n\big\}\Big) \\ &\leq P\big(\sigma_{max}^2 \min_{l \in \mathcal{N}} |\widehat{\Theta}_{ll}| \leq 1 - \frac{1}{2}\epsilon_n\big) + P\big(\mathcal{D}_1 \cap \big\{\sigma_{max}^2 \min_{l \in \mathcal{N}} |\widehat{\Theta}_{ll}| \geq 1 - \frac{1}{2}\epsilon_n\big\}\big) \\ &:= P_1 + P_2. \end{aligned}$$

It then suffices to bound P_1 and P_2 separately. First, P_1 can be bounded as

$$\begin{aligned} P_1 &= P\Big(\sigma_{max}^2 \min_{l \in \mathcal{N}} |\Theta_{ll} + \widehat{\Theta}_{ll} - \Theta_{ll}| \leq 1 - \frac{1}{2}\epsilon_n\Big) \\ &\leq P\Big(\sigma_{max}^2 \min_{l \in \mathcal{N}} \Theta_{ll} - \sigma_{max}^2 \max_{l \in \mathcal{N}} |\widehat{\Theta}_{ll} - \Theta_{ll}| \leq 1 - \frac{1}{2}\epsilon_n\Big) \\ &\leq P\Big(\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{max} \geq \frac{1}{2}\sigma_{max}^{-2}\epsilon_n\Big) \\ &= P\Big(\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{max} \geq b_1 \gamma^2 \tau^{1/2} \sqrt{\frac{\log p}{n}}\Big) \leq b_2 p^{2-\tau}, \end{aligned} \tag{9}$$

where the second inequality follows from the fact that $\min_{l \in \mathcal{N}} \Theta_{ll} = \min_{l \in \mathcal{N}} (\sigma_l^{-2} + \sum_{k \in \mathbf{ch}_l} \sigma_k^{-2} \beta_{kl}^2) \geq \min_{l \in \mathcal{N}} \sigma_l^{-2} \geq \sigma_{max}^{-2}$, and the last inequality follows from Lemma 3.

Next, to bound P_2, we have

$$\begin{aligned} P_2 \leq & P\big(\mathcal{D}_1 \cap \big\{\sigma_{max}^2 \min_{l \in \mathcal{N}} |\widehat{\Theta}_{ll}| \geq \frac{1}{2}\big\}\big) \\ \leq & P\Big(\Big\| \frac{\widehat{\mathbf{\Theta}}_{-ll}}{\widehat{\Theta}_{ll}} - \frac{\mathbf{\Theta}_{-ll}}{\widehat{\Theta}_{ll}} \Big\|_2 \geq 2b_1 \gamma^2 \tau^{1/2} \sqrt{\frac{d \log p}{L_2^2 n}}\ ,\ \sigma_{max}^2 \min_{l \in \mathcal{N}} |\widehat{\Theta}_{ll}| \geq \frac{1}{2}\Big) + \\ & P\Big(\Big\| \frac{\mathbf{\Theta}_{-ll}}{\widehat{\Theta}_{ll}} - \frac{\mathbf{\Theta}_{-ll}}{\Theta_{ll}} \Big\|_2 \geq 2b_1 \gamma^2 \tau^{1/2} \sqrt{\frac{d \log p}{L_2^2 n}}\ ,\ \sigma_{max}^2 \min_{l \in \mathcal{N}} |\widehat{\Theta}_{ll}| \geq \frac{1}{2}\Big) \\ := & P_{21} + P_{22} \end{aligned}$$

where the first inequality follows the fact that $\epsilon_n \leq 1$ for sufficiently large n.

To bound P_{21}, there holds

$$\begin{aligned}
P_{21} \leq & P\Big(\big\|\widehat{\mathbf{\Theta}}_{-ll} - \mathbf{\Theta}_{-ll}\big\|_2 \geq b_1\sigma_{max}^{-2}\gamma^2\tau^{1/2}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \\
\leq & P\Big(\big\|\widehat{\mathbf{\Theta}}_{-ll} - \mathbf{\Theta}_{-ll}\big\|_\infty \geq b_1\sigma_{max}^{-2}\gamma^2\tau^{1/2}\sqrt{\frac{\log p}{L_2^2 n}}\Big) \\
\leq & P\Big(\big\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\big\|_{max} \geq b_1\sigma_{max}^{-2}\gamma^2\tau^{1/2}\sqrt{\frac{\log p}{L_2^2 n}}\Big) \\
\leq & P\Big(\big\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\big\|_{max} \geq b_1\gamma^2\tau^{1/2}\sqrt{\frac{\log p}{n}}\Big) \leq b_2 p^{2-\tau}, \qquad (10)
\end{aligned}$$

where the second inequality follows from Theorem 1(b) in Ravikumar et al. (2011) which induces that $|\widehat{\mathbf{\Theta}}_{-ll} - \mathbf{\Theta}_{-ll}|$ has at most d non-zero elements and thus $\|\widehat{\mathbf{\Theta}}_{-ll} - \mathbf{\Theta}_{-ll}\|_2 \leq \sqrt{d}\|\widehat{\mathbf{\Theta}}_{-ll} - \mathbf{\Theta}_{-ll}\|_\infty$, and the last inequality follows from Lemma 3.

To bound P_{22}, we first note that $\Theta_{lk} = -\sigma_l^{-2}\beta_{lk} - \sigma_k^{-2}\beta_{kl} + \sum_{i\in \mathsf{ch}_l\cap \mathsf{ch}_k}\sigma_i^{-2}\beta_{il}\beta_{ik}$ for any $l \neq k$ and thus $\max_{l\neq k}|\Theta_{lk}| \leq \sigma_{min}^{-2}\beta_{max}(1 + d\beta_{max}) = \sigma_{min}^{-2}L_1$. Then, there holds

$$\begin{aligned}
P_{22} \leq & P\Big(\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{max} \geq 2b_1\frac{\min_{l\in\mathcal{N}}\Theta_{ll}\min_{l\in\mathcal{N}}|\widehat{\Theta}_{ll}|}{\max_{l\in\mathcal{N}}\|\mathbf{\Theta}_{-ll}\|_2}\gamma^2\tau^{1/2}\sqrt{\frac{d\log p}{L_2^2 n}}\ ,\ \sigma_{max}^2\min_{l\in\mathcal{N}}|\widehat{\Theta}_{ll}| \geq \frac{1}{2}\Big) \\
\leq & P\Big(\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{max} \geq 2b_1\frac{\min_{l\in\mathcal{N}}\Theta_{ll}\min_{l\in\mathcal{N}}|\widehat{\Theta}_{ll}|}{\max_{l\neq k}|\Theta_{lk}|}\gamma^2\tau^{1/2}\sqrt{\frac{\log p}{L_2^2 n}}\ ,\ \sigma_{max}^2\min_{l\in\mathcal{N}}|\widehat{\Theta}_{ll}| \geq \frac{1}{2}\Big) \\
\leq & P\Big(\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{max} \geq b_1\sigma_{max}^{-4}\sigma_{min}^2 L_1^{-1}\gamma^2\tau^{1/2}\sqrt{\frac{\log p}{L_2^2 n}}\Big) \\
\leq & P\Big(\big\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\big\|_{max} \geq b_1\gamma^2\tau^{1/2}\sqrt{\frac{\log p}{n}}\Big) \leq b_2 p^{2-\tau}, \qquad (11)
\end{aligned}$$

where the second inequality follows the fact that $\mathbf{\Theta}_{-ll}$ has at most d non-zero elements and thus $\|\mathbf{\Theta}_{-ll}\|_2 \leq \sqrt{d}\|\mathbf{\Theta}_{-ll}\|_\infty \leq \sqrt{d}\max_{l\neq k}|\Theta_{lk}|$, and the third inequality follows from the facts that $\min_{l\in\mathcal{N}}\Theta_{ll} \geq \sigma_{max}^{-2}$ and $\max_{l\neq k}|\Theta_{lk}| \leq \sigma_{min}^{-2}L_1$.

Therefore, combining (9), (10) and (11) yields that $P(\mathcal{D}_1) \leq 3b_2 p^{2-\tau}$. The desired result follows immediately with $a_2 = 3b_2$. ■

Appendix C. Proof for Theorem 2

Before providing necessary lemmas and proofs, we first give some notations. For each $l \in \mathcal{N}$, $e_{l,\mathcal{N}} = x_l + \mathbf{x}_{-l}^T\mathbf{\Theta}_{-ll}/\Theta_{ll} = \mathbf{x}^T\mathbf{\Theta}_{\cdot l}/\Theta_{ll}$, $\mathbf{e}_{l,\mathcal{N}} = \mathbf{X}\mathbf{\Theta}_{\cdot l}/\Theta_{ll} = (e_{1l,\mathcal{N}}, ..., e_{nl,\mathcal{N}})^T$, and $\widehat{\mathbf{e}}_{l,\mathcal{N}} = \mathbf{X}\widehat{\mathbf{\Theta}}_{\cdot l}/\widehat{\Theta}_{ll} = (\widehat{e}_{1l,\mathcal{N}}, ..., \widehat{e}_{nl,\mathcal{N}})^T$. Furthermore, for each $l \in \mathcal{N}$ and $k \in \mathcal{N}\backslash\{l\}$, we consider the following independence test statistic,

$$T(e_{l,\mathcal{N}}, x_k) = \frac{\mathrm{dcov}^2(e_{l,\mathcal{N}}, x_k)}{I_{lk,2}},$$

where $\text{dcov}^2(e_{l,\mathcal{N}}, x_k) = I_{lk,1} + I_{lk,2} - 2I_{lk,3}$, $I_{lk,1} = E[|e_{l,\mathcal{N}} - e'_{l,\mathcal{N}}||x_k - x'_k|]$, $I_{lk,2} = E[|e_{l,\mathcal{N}} - e'_{l,\mathcal{N}}|]E[|x_k - x'_k|]$ and $I_{lk,3} = E\big[E[|e_{l,\mathcal{N}} - e'_{l,\mathcal{N}}||e_{l,\mathcal{N}}]E[|x_k - x'_k||x_k]\big]$ with $e'_{l,\mathcal{N}}$ and x'_k denoting an independent copy of $e_{l,\mathcal{N}}$ and x_k, respectively. Its sample and estimated versions are

$$\widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k) = \frac{\widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)}{\widehat{I}_{lk,2}} \text{ and } \widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) = \frac{\widehat{\text{dcov}}^2(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k)}{\widetilde{I}_{lk,2}}.$$

Here, $\widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k) = \widehat{I}_{lk,1} + \widehat{I}_{lk,2} - 2\widehat{I}_{lk,3}$ with $\widehat{I}_{lk,1} = \frac{1}{n^2}\sum_{i,j=1}^n |e_{il,\mathcal{N}} - e_{jl,\mathcal{N}}||x_{ik} - x_{jk}|$, $\widehat{I}_{lk,2} = (\frac{1}{n^2}\sum_{i,j=1}^n |e_{il,\mathcal{N}} - e_{jl,\mathcal{N}}|)(\frac{1}{n^2}\sum_{i,j=1}^n |x_{ik} - x_{jk}|)$, $\widehat{I}_{lk,3} = \frac{1}{n^3}\sum_{i,j,h=1}^n |e_{il,\mathcal{N}} - e_{hl,\mathcal{N}}||x_{jk} - x_{hk}|$, and $\widehat{\text{dcov}}^2(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) = \widetilde{I}_{lk,1} + \widetilde{I}_{lk,2} - 2\widetilde{I}_{lk,3}$ with $\widetilde{I}_{lk,1} = \frac{1}{n^2}\sum_{i,j=1}^n |\widehat{e}_{il,\mathcal{N}} - \widehat{e}_{jl,\mathcal{N}}||x_{ik} - x_{jk}|$, $\widetilde{I}_{lk,2} = (\frac{1}{n^2}\sum_{i,j=1}^n |\widehat{e}_{il,\mathcal{N}} - \widehat{e}_{jl,\mathcal{N}}|)(\frac{1}{n^2}\sum_{i,j} |x_{ik} - x_{jk}|)$, $\widetilde{I}_{lk,3} = \frac{1}{n^3}\sum_{i,j,h=1}^n |\widehat{e}_{il,\mathcal{N}} - \widehat{e}_{hl,\mathcal{N}}||x_{jk} - x_{hk}|$.

Lemma 4 *Suppose that all the assumptions in Lemma 2 and Assumption 3 are satisfied. For any $l \in \mathcal{N}$, with probability at least $1 - a_2 p^{2-\tau}$, there holds*

$$\frac{1}{\sqrt{n}}\|\widehat{\mathbf{e}}_{l,\mathcal{N}} - \mathbf{e}_{l,\mathcal{N}}\|_2 \leq a_1\gamma^2(\tau\lambda_{max})^{1/2}\sqrt{\frac{d\log p}{L_2^2 n}}.$$

Proof of Lemma 4: Recall that $\mathbf{e}_{l,\mathcal{N}} = \mathbf{X}\,\boldsymbol{\Theta}_{\cdot l}/\Theta_{ll}$ and $\widehat{\mathbf{e}}_{l,\mathcal{N}} = \mathbf{X}\,\widehat{\boldsymbol{\Theta}}_{\cdot l}/\widehat{\Theta}_{ll}$, and thus

$$\widehat{\mathbf{e}}_{l,\mathcal{N}} - \mathbf{e}_{l,\mathcal{N}} = \mathbf{X}\Big(\frac{\widehat{\boldsymbol{\Theta}}_{\cdot l}}{\widehat{\Theta}_{ll}} - \frac{\boldsymbol{\Theta}_{\cdot l}}{\Theta_{ll}}\Big).$$

Furthermore, it follows from Assumption 3 that

$$\frac{1}{\sqrt{n}}\|\widehat{\mathbf{e}}_{l,\mathcal{N}} - \mathbf{e}_{l,\mathcal{N}}\|_2 \leq \frac{1}{\sqrt{n}}\|\,\mathbf{X}\,\|_2\,\Big\|\frac{\widehat{\boldsymbol{\Theta}}_{\cdot l}}{\widehat{\Theta}_{ll}} - \frac{\boldsymbol{\Theta}_{\cdot l}}{\Theta_{ll}}\Big\|_2 \leq (\lambda_{max})^{1/2}\Big\|\frac{\widehat{\boldsymbol{\Theta}}_{\cdot l}}{\widehat{\Theta}_{ll}} - \frac{\boldsymbol{\Theta}_{\cdot l}}{\Theta_{ll}}\Big\|_2.$$

It further implies that

$$P\Big(\frac{1}{\sqrt{n}}\|\widehat{\mathbf{e}}_{l,\mathcal{N}} - \mathbf{e}_{l,\mathcal{N}}\|_2 \geq a_1\gamma^2(\tau\lambda_{max})^{1/2}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \leq P\Big(\Big\|\frac{\widehat{\boldsymbol{\Theta}}_{\cdot l}}{\widehat{\Theta}_{ll}} - \frac{\boldsymbol{\Theta}_{\cdot l}}{\Theta_{ll}}\Big\|_2 \geq a_1\gamma^2\tau^{1/2}\sqrt{\frac{d\log p}{L_2^2 n}}\Big),$$

and then by Lemma 2, there holds

$$P\Big(\frac{1}{\sqrt{n}}\|\widehat{\mathbf{e}}_{l,\mathcal{N}} - \mathbf{e}_{l,\mathcal{N}}\|_2 \geq a_1\gamma^2(\tau\lambda_{max})^{1/2}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \leq a_2 p^{2-\tau}.$$

This completes the proof. ■

Lemma 5 *Suppose that all the assumptions in Lemma 4 are satisfied, and $n = \Omega(d^6 \log p)$. Then, for any $l \in \mathcal{N}$ and $k \in \mathcal{N}\backslash\{l\}$, there exist some positive constants b_3 and b_4 such that with probability at least $1 - b_4 p^{2-\tau}$, there holds*

$$\big|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - \widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)\big| \leq b_3\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}},$$

where $L_3 = \max\{1, 16\lambda_{max}(dL_1^2(\frac{\sigma_{max}}{\sigma_{min}})^4 + 1)^{1/2}\}$.

Proof of Lemma 5: Note that

$$\begin{aligned}
|\widetilde{I}_{lk,1} - \widehat{I}_{lk,1}| &\leq \frac{1}{n^2}\sum_{i,j=1}^{n}|x_{ik}-x_{jk}|\Big||\widehat{e}_{il,\mathcal{N}}-\widehat{e}_{jl,\mathcal{N}}|-|e_{il,\mathcal{N}}-e_{jl,\mathcal{N}}|\Big| \\
&\leq \frac{1}{n^2}\sum_{i,j=1}^{n}|x_{ik}-x_{jk}|\Big(|\widehat{e}_{il,\mathcal{N}}-e_{il,\mathcal{N}}|+|\widehat{e}_{jl,\mathcal{N}}-e_{jl,\mathcal{N}}|\Big) \\
&= \frac{2}{n^2}\sum_{i,j=1}^{n}|x_{ik}-x_{jk}||\widehat{e}_{il,\mathcal{N}}-e_{il,\mathcal{N}}| \\
&\leq \frac{2}{n}\sum_{i=1}^{n}|x_{ik}||\widehat{e}_{il,\mathcal{N}}-e_{il,\mathcal{N}}| + 2\Big(\frac{1}{n}\sum_{i=1}^{n}|x_{ik}|\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}|\widehat{e}_{il,\mathcal{N}}-e_{il,\mathcal{N}}|\Big) \\
&\leq 4\frac{1}{\sqrt{n}}\|\mathbf{x}_k\|_2\frac{1}{\sqrt{n}}\|\widehat{\mathbf{e}}_{l,\mathcal{N}}-\mathbf{e}_{l,\mathcal{N}}\|_2 \\
&\leq 4(\lambda_{max})^{1/2}\frac{1}{\sqrt{n}}\|\widehat{\mathbf{e}}_{l,\mathcal{N}}-\mathbf{e}_{l,\mathcal{N}}\|_2,
\end{aligned}$$

where the second last inequality follows from Hölder 's inequality and Jensen's inequality, and the last inequality follows from the fact that $\frac{1}{n}\|\mathbf{x}_k\|_2^2 \leq \Lambda_{max}\Big(\frac{1}{n}\mathbf{X}^T\mathbf{X}\Big)$ and Assumption 3.

Then, it follows from Lemma 4 that

$$\begin{aligned}
&P\Big(|\widetilde{I}_{lk,1}-\widehat{I}_{lk,1}| \geq 4a_1\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \\
\leq &P\Big(\frac{1}{\sqrt{n}}\|\widehat{\mathbf{e}}_{l,\mathcal{N}}-\mathbf{e}_{l,\mathcal{N}}\|_2 \geq a_1\gamma^2(\tau\lambda_{max})^{1/2}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \leq a_2 p^{2-\tau}. \qquad (12)
\end{aligned}$$

Following a similar treatment, there also holds

$$P\Big(|\widetilde{I}_{lk,2}-\widehat{I}_{lk,2}| \geq 4a_1\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \leq a_2 p^{2-\tau}, \qquad (13)$$

$$P\Big(|\widetilde{I}_{lk,3}-\widehat{I}_{lk,3}| \geq 4a_1\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \leq a_2 p^{2-\tau}. \qquad (14)$$

Recall that $\widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}},\mathbf{x}_k) = \widehat{I}_{lk,1}+\widehat{I}_{lk,2}-2\widehat{I}_{lk,3}$ and $\widehat{\text{dcov}}^2(\widehat{\mathbf{e}}_{l,\mathcal{N}},\mathbf{x}_k) = \widetilde{I}_{lk,1}+\widetilde{I}_{lk,2}-2\widetilde{I}_{lk,3}$, and thus we have

$$\big|\widehat{\text{dcov}}^2(\widehat{\mathbf{e}}_{l,\mathcal{N}},\mathbf{x}_k) - \widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}},\mathbf{x}_k)\big| \leq |\widetilde{I}_{lk,1}-\widehat{I}_{lk,1}| + |\widetilde{I}_{lk,2}-\widehat{I}_{lk,2}| + 2|\widetilde{I}_{lk,3}-\widehat{I}_{lk,3}|.$$

Therefore, combining (12), (13) and (14) yields that

$$P\Big(\big|\widehat{\text{dcov}}^2(\widehat{\mathbf{e}}_{l,\mathcal{N}},\mathbf{x}_k) - \widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}},\mathbf{x}_k)\big| \geq 16a_1\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \leq 3a_2 p^{2-\tau}. \qquad (15)$$

Next, recall that $\mathbf{e}_{l,\mathcal{N}} = \mathbf{X}\,\boldsymbol{\Theta}_{\cdot l}/\Theta_{ll}$, then there holds

$$\frac{1}{\sqrt{n}}\|\mathbf{e}_{l,\mathcal{N}}\|_2 \leq \frac{1}{\sqrt{n}}\|\,\mathbf{X}\,\|_2\,\|\boldsymbol{\Theta}_{\cdot l}/\Theta_{ll}\|_2 \leq (\lambda_{max})^{1/2}(dL_1^2\Big(\frac{\sigma_{max}}{\sigma_{min}}\Big)^4+1)^{1/2},$$

where the last inequality follows from Assumptions 3 and the fact that $\min_{l\in\mathcal{N}}\Theta_{ll} \geq \sigma_{max}^{-2}$ and $\max_{l\neq k}|\Theta_{lk}| \leq \sigma_{min}^{-2}\beta_{max}(1+d\beta_{max}) = \sigma_{min}^{-2}L_1$. Then, we have

$$\begin{aligned}
|\widehat{I}_{lk,1}| &= \frac{1}{n^2}\sum_{i,j=1}^{n}|x_{ik}-x_{jk}||e_{il,\mathcal{N}}-e_{jl,\mathcal{N}}| \leq \frac{1}{n^2}\sum_{i,j=1}^{n}(|x_{ik}|+|x_{jk}|)(|e_{il,\mathcal{N}}|+|e_{jl,\mathcal{N}}|)\\
&= \frac{2}{n}\sum_{i=1}^{n}|x_{ik}||e_{il,\mathcal{N}}| + 2\Big(\frac{1}{n}\sum_{i=1}^{n}|x_{ik}|\Big)\Big(\frac{1}{n}\sum_{j=1}^{n}|e_{jl,\mathcal{N}}|\Big) \leq 4(\lambda_{max})^{1/2}\frac{1}{\sqrt{n}}\|\mathbf{e}_{l,\mathcal{N}}\|_2\\
&\leq 4\lambda_{max}(dL_1^2\Big(\frac{\sigma_{max}}{\sigma_{min}}\Big)^4+1)^{1/2}.
\end{aligned}$$

where the second inequality follows from Hölder 's inequality, Jensen's inequality and Assumption 3 with the fact that $\frac{1}{n}\|\mathbf{x}_k\|_2^2 \leq \Lambda_{max}\Big(\frac{1}{n}\mathbf{X}^T\mathbf{X}\Big)$. Following a similar treatment, there also holds

$$|\widehat{I}_{lk,2}| \leq 4\lambda_{max}(dL_1^2\Big(\frac{\sigma_{max}}{\sigma_{min}}\Big)^4+1)^{1/2} \text{ and } |\widehat{I}_{lk,3}| \leq 4\lambda_{max}(dL_1^2\Big(\frac{\sigma_{max}}{\sigma_{min}}\Big)^4+1)^{1/2},$$

Therefore, we obtain that

$$|\widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}},\mathbf{x}_k)| \leq |\widehat{I}_{lk,1}| + |\widehat{I}_{lk,2}| + 2|\widehat{I}_{lk,3}| \leq 16\lambda_{max}(dL_1^2\Big(\frac{\sigma_{max}}{\sigma_{min}}\Big)^4+1)^{1/2}. \tag{16}$$

Note that it follows from (B.10) in Li et al. (2012) that $\widehat{I}_{lk,2}$ converges to $I_{lk,2}$ with high probability, and $I_{lk,2}$ must be a positive constant by its definition. Thus, when n is sufficiently large, there holds true that $\min_{l\neq k}\widehat{I}_{lk,2} \geq b_5$ with $b_5 = I_{lk,2}/2$. Then, we denote

$$\xi_n = 8a_1/b_5\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}}$$

and define the event

$$\mathcal{D}_2 = \Big\{\big|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}},\mathbf{x}_k) - \widehat{T}(\mathbf{e}_{l,\mathcal{N}},\mathbf{x}_k)\big| \geq b_3\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}}\Big\},$$

where $b_3 = 16a_1/b_6$ with $b_6 = b_5\min\{1/4, b_5\}$.

Then, $P(\mathcal{D}_2)$ can be decomposed as

$$\begin{aligned}
P(\mathcal{D}_2) &= P\Big(\mathcal{D}_2 \cap \{\widetilde{I}_{lk,2} \leq b_5 - \frac{1}{2}b_5\xi_n\}\Big) + P\Big(\mathcal{D}_2 \cap \{\widetilde{I}_{lk,2} \geq b_5 - \frac{1}{2}b_5\xi_n\}\Big)\\
&\leq P\Big(\widetilde{I}_{lk,2} \leq b_5 - \frac{1}{2}b_5\xi_n\Big) + P\Big(\mathcal{D}_2 \cap \{\widetilde{I}_{lk,2} \geq b_5 - \frac{1}{2}b_5\xi_n\}\Big)\\
&= P_3 + P_4.
\end{aligned}$$

To bound P_3, we notice that

$$\begin{aligned}
P_3 &\leq P\Big(\min_{l\neq k}\widehat{I}_{lk,2} - |\widetilde{I}_{lk,2} - \widehat{I}_{lk,2}| \leq b_5 - \frac{1}{2}b_5\xi_n\Big)\\
&\leq P\Big(|\widetilde{I}_{lk,2} - \widehat{I}_{lk,2}| \geq \frac{1}{2}b_5\xi_n\Big)\\
&\leq P\Big(|\widetilde{I}_{lk,2} - \widehat{I}_{lk,2}| \geq 4a_1\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{d\log p}{L_2^2 n}}\Big) \leq a_2 p^{2-\tau},
\end{aligned} \tag{17}$$

where the second inequality follows from the fact that $\min_{l\neq k}\widehat{I}_{lk,2} \geq b_5$, and the last inequality follows from (13).

Next, we turn to bound P_4. Note that when n is sufficiently large, $\xi_n \in (0,1)$. Thus,

$$\begin{aligned}
P_4 \leq& P\Big(\mathcal{D}_2 \cap \{\widetilde{I}_{lk,2} \geq \frac{1}{2}b_5\}\Big)\\
\leq& P\Big(\Big|\frac{\widehat{\text{dcov}}^2(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k)}{\widetilde{I}_{lk,2}} - \frac{\widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)}{\widetilde{I}_{lk,2}}\Big| \geq \frac{1}{2}b_3\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}}\,,\ \widetilde{I}_{lk,2} \geq \frac{1}{2}b_5\Big) +\\
& P\Big(\Big|\frac{\widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)}{\widetilde{I}_{lk,2}} - \frac{\widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)}{\widehat{I}_{lk,2}}\Big| \geq \frac{1}{2}b_3\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}}\,,\ \widetilde{I}_{lk,2} \geq \frac{1}{2}b_5\Big)\\
:=& P_{41} + P_{42}.
\end{aligned}$$

For the first term P_{41}, there holds

$$\begin{aligned}
P_{41} \leq& P\Big(|\widehat{\text{dcov}}^2(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - \widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)| \geq \frac{1}{4}b_3b_5\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}}\Big)\\
\leq& P\Big(\big|\widehat{\text{dcov}}^2(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - \widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)\big| \geq 16a_1\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{d\log p}{L_2^2 n}}\Big)\\
\leq& 3a_2 p^{2-\tau}.
\end{aligned} \tag{18}$$

For the second term P_{42}, there holds

$$\begin{aligned}
P_{42} \leq& P\Big(|\widetilde{I}_{lk,2} - \widehat{I}_{lk,2}| \geq \frac{1}{2}b_3\frac{\widetilde{I}_{lk,2}\widehat{I}_{lk,2}}{|\widehat{\text{dcov}}^2(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)|}\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}}\,,\ \widetilde{I}_{lk,2} \geq \frac{1}{2}b_5\Big)\\
\leq& P\Big(|\widetilde{I}_{lk,2} - \widehat{I}_{lk,2}| \geq \frac{1}{4}b_3\frac{b_5^2}{16\lambda_{max}(dL_1^2\big(\frac{\sigma_{max}}{\sigma_{min}}\big)^4 + 1)^{1/2}}\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}}\Big)\\
\leq& P\Big(|\widetilde{I}_{lk,2} - \widehat{I}_{lk,2}| \geq 4a_1\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{d\log p}{L_2^2 n}}\Big)\\
\leq& a_2 p^{2-\tau},
\end{aligned} \tag{19}$$

where the second inequality follows the facts that $\min_{l\neq k}\widehat{I}_{lk,2} \geq b_5$ and (16).

Therefore, combining (17), (18) and (19), we have

$$P(\mathcal{D}_2) = P\Big(|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - \widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)| \geq b_3\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d\log p}{L_2^2 n}}\Big) \leq 5a_2 p^{2-\tau}.$$

This completes the proof. ■

Lemma 6 *Suppose that Assumption 2 is satisfied. For any $l \in \mathcal{N}$ and $k \in \mathcal{N}\backslash\{l\}$, there exists some positive constants b_7, b_8, b_9 and $\eta < 1/2$ such that*

$$P\Big(|\widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k) - T(e_{l,\mathcal{N}}, x_k)| \geq b_7 n^{-\eta}\Big) \leq b_8\exp(-b_9 n^{(1-2\eta)/3}).$$

Lemma 6 ensures that $\widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)$ converges to $T(e_{l,\mathcal{N}}, x_k)$ with high probability. Its proof is similar to Theorem 1 in Li et al. (2012), and thus omitted here.

Proof of Theorem 2: Note that the proposed method needs to conduct $p(p-1)$ times of independence tests to estimate $\mathcal{A}_0$. Specifically, for any $l \in \mathcal{N}$ and $k \in \mathcal{N}\backslash\{l\}$, we consider the following testing hypothesis,

$$\mathcal{H}_{lk,0} : e_{l,\mathcal{N}} \perp\!\!\!\perp x_k \ \text{ vs } \ \mathcal{H}_{lk,1} : e_{l,\mathcal{N}} \not\perp\!\!\!\perp x_k,$$

and denote $\mathcal{E}_{lk} = \mathcal{E}_{lk}^I \cup \mathcal{E}_{lk}^{II}$, where

$$\text{Type I error} : \mathcal{E}_{lk}^I = \Big\{ n\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) > (\Phi^{-1}(1-\alpha/2))^2 \text{ and } T(e_{l,\mathcal{N}}, x_k) = 0\Big\},$$
$$\text{Type II error} : \mathcal{E}_{lk}^{II} = \Big\{ n\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) \leq (\Phi^{-1}(1-\alpha/2))^2 \text{ and } T(e_{l,\mathcal{N}}, x_k) > 0\Big\}.$$

It is clear that $P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0) \leq P\big(\cup_{l\neq k}\mathcal{E}_{lk}\big) \leq p^2\sup_{l\neq k} P\big(\mathcal{E}_{lk}\big)$, and thus we turn to bound $P\big(\mathcal{E}_{lk}\big)$.

By Assumption 2, it can be verfied that x_k and $e_{l,\mathcal{N}}$ are also sub-Gaussian distributed, and thus there must exist some positive constants b_{10} and b_{11} such that $\max_{k\in\mathcal{N}} E|x_k| \leq b_{10}$ and $\max_{l\in\mathcal{N}} E|e_{l,\mathcal{N}}| \leq b_{11}$. Then, we set the significance level as $\alpha := \alpha_n = 2(1-\Phi(a_6\sqrt{n}\rho_{n,0}))$ with $a_6 = \frac{1}{\sqrt{8b_{10}b_{11}}}$, and it is clear that $\alpha_n \to 0$ as $n \to \infty$. Thus, there holds

$$\begin{aligned}
\sup_{l\neq k} P(\mathcal{E}_{lk}^I) &= \sup_{l\neq k} P\Big(n\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) > (\Phi^{-1}(1-\alpha/2))^2\Big) \\
&= \sup_{l\neq k} P\Big(|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - T(e_{l,\mathcal{N}}, x_k)| > \frac{1}{n}(\Phi^{-1}(1-\alpha/2))^2\Big) \\
&= \sup_{l\neq k} P\Big(|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - T(e_{l,\mathcal{N}}, x_k)| > \frac{\rho_{n,0}^2}{8b_{10}b_{11}}\Big). \qquad (20)
\end{aligned}$$

By Assumption 4 and the fact that $I_{lk,2} \leq 4E|x_k|E|e_{l,\mathcal{N}}| \leq 4b_{10}b_{11}$, we have $T(e_{l,\mathcal{N}}, x_k) = \text{dcov}^2(e_{l,\mathcal{N}}, x_k)/I_{lk,2} \geq \frac{\rho_{n,0}^2}{4b_{10}b_{11}}$ if $l \notin \mathcal{A}_0$, and then

$$\begin{aligned}
\sup_{l\neq k} P(\mathcal{E}_{lk}^{II}) &= \sup_{l\neq k} P\Big(n\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) \leq (\Phi^{-1}(1-\alpha/2))^2\Big) \\
&\leq \sup_{l\neq k} P\Big(|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - T(e_{l,\mathcal{N}}, x_k)| \geq \frac{\rho_{n,0}^2}{4b_{10}b_{11}} - \frac{1}{n}(\Phi^{-1}(1-\alpha/2))^2\Big) \\
&= \sup_{l\neq k} P\Big(|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - T(e_{l,\mathcal{N}}, x_k)| \geq \frac{\rho_{n,0}^2}{8b_{10}b_{11}}\Big). \qquad (21)
\end{aligned}$$

Note that when n is sufficiently large, there must exist some positive constant b_{12} such that $b_{12}\sqrt{\frac{d^6 \log p}{n}} \geq \sqrt{\frac{L_3^2 d \log p}{L_2^2 n}}$, and then by Assumption 4, we have $\frac{\rho_{n,0}^2}{8b_{10}b_{11}} \geq b_3 b_{12}\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{d^6 \log p}{n}} + b_7 n^{-\eta} \geq b_3\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d \log p}{L_2^2 n}} + b_7 n^{-\eta}$.

It then follows from (20), (21), Lemmas 5 and 6 that

$$\begin{aligned}
\sup_{l\neq k} P(\mathcal{E}_{lk}) \leq& \sup_{l\neq k} P(\mathcal{E}_{lk}^{I}) + \sup_{l\neq k} P(\mathcal{E}_{lk}^{II}) \\
\leq& 2\sup_{l\neq k} P\Big(|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - T(e_{l,\mathcal{N}}, x_k)| \geq \frac{\rho_{n,0}^2}{8b_{10}b_{11}}\Big) \\
\leq& 2\sup_{l\neq k} P\Big(|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - \widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)| \geq b_3\gamma^2\tau^{1/2}\lambda_{max}\sqrt{\frac{L_3^2 d \log p}{L_2^2 n}}\Big) + \\
& 2\sup_{l\neq k} P\Big(|\widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k) - T(e_{l,\mathcal{N}}, x_k)| \geq b_7 n^{-\eta}\Big) \\
\leq& 2b_4 p^{2-\tau} + 2b_8 \exp\{-b_9 n^{(1-2\eta)/3}\}.
\end{aligned}$$

Hence, we have

$$P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0) \leq P\big(\cup_{l\neq k}\mathcal{E}_{lk}\big) \leq p^2 \sup_{l\neq k} P\big(\mathcal{E}_{lk}\big) \leq 2b_4 p^{4-\tau} + 2b_8 p^2 \exp\{-b_9 n^{(1-2\eta)/3}\}.$$

This completes the proof of Theorem 2. ■

Appendix D. Proof for Theorem 3

Proof of Theorem 3: Note that

$$\begin{aligned}
P(\widehat{\mathcal{G}} \neq \mathcal{G}) =& P\Big((\cup_{t=0}^{T-1}\{\widehat{\mathcal{A}}_t \neq \mathcal{A}_t\}) \cup \{\widehat{\mathcal{E}} \neq \mathcal{E}\}\Big) \\
\leq& P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0) + \sum_{t=1}^{T-1} P(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t | \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}) + \\
& P(\widehat{\mathcal{E}} \neq \mathcal{E} | \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{T-1} = \mathcal{A}_{T-1}). \qquad (22)
\end{aligned}$$

By Theorem 2, we have $P(\widehat{\mathcal{A}}_0 = \mathcal{A}_0) \geq 1 - a_3 p^{4-\tau} - a_4 p^2 \exp\{-a_5 n^{(1-2\eta)/3}\}$. Then, it suffices to bound $P\Big(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t \Big| \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\Big)$ and $P(\widehat{\mathcal{E}} \neq \mathcal{E} | \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{T-1} = \mathcal{A}_{T-1})$, separately.

Note that $|\mathcal{S}_t|$ can be diverging with n or just a fixed constant, and thus the proof of Theorem 2 is not applicable to $\mathcal{S}_t$ directly. Instead, for any $t = 1, 2, ..., T-1$, we replace the term p^τ in Theorem 1 of Ravikumar et al. (2011) by $|\mathcal{S}_t|^4 n^{\tau-4}$, and then it is can be verified that all the required conditions in the proof of Theorem 1 of Ravikumar et al. (2011) are still satisfied if $n = \Omega\Big(T^{1/(\tau-4)}(d^6 + \beta_{min}^{-2})(\log(\max\{p, n\}))^{3/(1-2\eta)}\Big)$. Thus, there must exist some positive constants b_{13} and b_{14} such that with probability at least $1 - b_{14}|\mathcal{S}_t|^{-2} n^{4-\tau}$, there holds

$$\|\widehat{\mathbf{\Theta}}^{\mathcal{S}_t} - \mathbf{\Theta}^{\mathcal{S}_t}\|_{max} \leq b_{13}\gamma^2 \tau^{1/2} \sqrt{\frac{\log(\max\{|\mathcal{S}_t|, n\})}{n}},$$

provided that $\lambda_{n,t} \propto \sqrt{\log(\max\{|\mathcal{S}_t|, n\})/n}$. Then, following a similar treatment as the proof of Theorem 2 and Corollary 3, there must exist some positive constants b_{15}, b_{16}, b_{17} and b_{18} such that

$$P\Big(\widehat{\mathcal{A}}_t = \mathcal{A}_t \Big| \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\Big) \geq 1 - b_{15} n^{4-\tau} - b_{16}|\mathcal{S}_t|^2 \exp\{-b_{17} n^{(1-2\eta)/3}\}, \quad (23)$$

$$P\Big(\big\{\widehat{\text{pa}}_l = \text{pa}_l : l \in \widehat{\mathcal{A}}_t\big\} \Big| \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_t = \mathcal{A}_t\Big) \geq 1 - b_{18}|\mathcal{S}_t|^{-2} n^{4-\tau}. \quad (24)$$

Furthermore, to bound $P(\widehat{\mathcal{E}} \neq \mathcal{E} | \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{T-1} = \mathcal{A}_{T-1})$, we notice that

$$\begin{aligned} P\Big(\widehat{\mathcal{E}} \neq \mathcal{E} | \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{T-1} = \mathcal{A}_{T-1}\Big) &\leq \sum_{t=0}^{T-2} P\Big(\big\{\widehat{\text{pa}}_l \neq \text{pa}_l : l \in \widehat{\mathcal{A}}_t\big\} \Big| \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_t = \mathcal{A}_t\Big) \\ &\leq a_7 p^{2-\tau} + b_{18}(T-2)|\mathcal{S}_t|^{-2} n^{4-\tau}, \end{aligned} \quad (25)$$

where the last inequality follows from (24) and Corollary 3.

Finally, combing (23) and (25) implies that

$$\begin{aligned} &P(\widehat{\mathcal{G}} \neq \mathcal{G}) \\ &\leq a_3 p^{4-\tau} + a_4 p^2 \exp\{-a_5 n^{(1-2\eta)/3}\} + (T-1)\Big(b_{15} n^{4-\tau} + b_{16}|\mathcal{S}_t|^2 \exp\{-b_{17} n^{(1-2\eta)/3}\}\Big) \\ &\qquad + a_7 p^{2-\tau} + b_{18}(T-2)|\mathcal{S}_t|^{-2} n^{4-\tau} \\ &\leq a_3 p^{4-\tau} + a_7 p^{2-\tau} + b_{19} T n^{4-\tau} + b_{20} T p^2 \exp\{-b_{21} n^{(1-2\eta)/3}\}, \end{aligned}$$

where $b_{19} = 2\max\{b_{15}, b_{18}\}$, $b_{20} = \max\{a_4, b_{16}\}$ and $b_{21} = \min\{a_5, b_{17}\}$. Thus, when $n = \Omega\Big(T^{1/(\tau-4)}(d^6 + \beta_{min}^{-2})(\log(\max\{p, n\}))^{3/(1-2\eta)}\Big)$, it is clear that $P(\widehat{\mathcal{G}} \neq \mathcal{G}) \to 0$ as n diverges. This completes the proof of Theorem 3. ■

Appendix E. Proof for Corollary 4

Lemma 7 *Suppose that Assumption 1 holds and for any $j \in \mathcal{N}$, ϵ_j/σ_j has $(4m)$-th bounded moment. Then there exist some positive constants b_{22}, b_{23} and $4 < \tau < 4+m$ such that with probability at least $1 - b_{23} p^{2-\tau}$, there holds*

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{max} \leq b_{22} \sqrt{\frac{p^{\tau/m}}{n}},$$

provided that $n = \Omega(d^2 p^{\tau/m})$ *and* $\lambda_{n,0} \propto \sqrt{\frac{p^{\tau/m}}{n}}$.

Note that the assumption that for any $j \in \mathcal{N}$, ϵ_j/σ_j has $(4m)$-th bounded moment, implies that $x_j/\sqrt{\Sigma_{jj}}$ also has $4m$-th bounded moment, and the detailed proof is provided in the proof of Theorem 3 in Appendix of Ghoshal and Honorio (2018). Clearly, the condition required in Lemma 2 of Ravikumar et al. (2011) is satisfied, and Lemma 7 is a standard consistency result by Lemma 2 of Ravikumar et al. (2011) for the precision matrix estimated by with graphical Lasso. Thus, we omit the proof of Lemma 7 here.

Lemma 8 *Suppose that all the assumptions in Lemma 7 and Assumption 3 are satisfied, and* $n = \Omega(d^6 p^{\tau/m})$. *Then, for any* $l \in \mathcal{N}$ *and* $k \in \mathcal{N}\backslash\{l\}$, *there exist some positive constants* b_{24} *and* b_{25} *such that with probability at least* $1 - b_{25}p^{2-\tau}$, *there holds*

$$\left|\widehat{T}(\widehat{\mathbf{e}}_{l,\mathcal{N}}, \mathbf{x}_k) - \widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k)\right| \leq b_{24}\lambda_{max}\sqrt{\frac{L_3^2 d p^{\tau/m}}{L_2^2 n}}.$$

The proof of Lemma 8 can be established as that in Lemma 5 with slight modification, and thus we omit it here.

Lemma 9 *Suppose that for any* $j \in \mathcal{N}$, ϵ_j/σ_j *has* $(4m)$*-th bounded moment. Then for any* $l \in \mathcal{N}$ *and* $k \in \mathcal{N}\backslash\{l\}$, *there exist some positive constants* b_{26}, b_{27}, $\eta < \frac{1}{2}$ *and* $\frac{1}{2m} < \phi < \frac{1}{2} - \eta$ *such that*

$$P\Big(\left|\widehat{T}(\mathbf{e}_{l,\mathcal{N}}, \mathbf{x}_k) - T(e_{l,\mathcal{N}}, x_k)\right| \geq b_{26}n^{-\eta}\Big) \leq b_{27}n^{1-2m\phi}.$$

Proof of Lemma 9: Note that the proof of Lemma 9 is similar as that of Theorem 1 in Li et al. (2012) except that we need to replace the upper bounds in (B.5), (B.13), (A.15) of Li et al. (2012) and modify all the relevant parts correspondingly. Specifically, to replace the upper bound in (B.5), we notice that for $0 < \phi < \frac{1}{2} - \eta$ and some positive constant b, there holds

$$P(|x_{ik}| \geq bn^{\phi/2}/2) \leq \frac{E(|x_k|^{4m})}{b^{4m}n^{2m\phi}},$$

where the inequality follows from Chebyshev's inequality and the fact that x_j also has $(4m)$-th bounded moment with mean zero. Thus, the upper bound in (B.5) of Theorem 1 in Li et al. (2012) can be modified as

$$P(|\widehat{I}^*_{lk,12}| > \epsilon/2) \leq b_{28}n^{1-2m\phi},$$

where $\widehat{I}^*_{lk,12} = \frac{1}{n(n-1)}\sum_{i\neq j}|e_{il,\mathcal{N}} - e_{jl,\mathcal{N}}||x_{ik} - x_{jk}|1(|e_{il,\mathcal{N}} - e_{jl,\mathcal{N}}||x_{ik} - x_{jk}| > b^2n^{\phi})$ with $1(\cdot)$ denoting the indicator function and b_{28} is a positive constant. Then, we use similar argument as above to replace (B.13) and (A.15) in Theorem 1 of Li et al. (2012) and also modify all the relevant parts correspondingly. This completes the proof. ■

Proof of Corollary 4: The proof of Corollary 4 can be completed following similar arguments as the proof in Theorem 3. Specifically, $P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0)$ and $P\Big(\big\{\widehat{\text{pa}}_l \neq \text{pa}_l : l \in \widehat{\mathcal{A}}_0\big\}\Big|\widehat{\mathcal{A}}_0 = \mathcal{A}_0\Big)$ can

be bounded by Lemmas 8 and 9, and following a similar treatment as in the proof of Theorem 2 and Corollary 3. Thus, there exist some positive constants b_{29}, b_{30} and b_{31} such that

$$P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0) \leq b_{29}p^{4-\tau} + b_{30}p^2 n^{1-2m\phi}, \tag{26}$$

$$P\Big(\{\widehat{\text{pa}}_l \neq \text{pa}_l : l \in \widehat{\mathcal{A}}_0\}\Big|\widehat{\mathcal{A}}_0 = \mathcal{A}_0\Big) \leq b_{31}p^{2-\tau}. \tag{27}$$

To bound $P\Big(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t\Big|\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\Big)$, we also replace p^τ in Theorem 1 of Ravikumar et al. (2011) by $|\mathcal{S}_t|^4 n^{\tau-4}$. Then, there must exist some positive constant b_{32} and b_{33} such that with probability at least $1 - b_{33}|\mathcal{S}_t|^{-2}n^{4-\tau}$, there holds

$$\|\widehat{\mathbf{\Theta}}^{\mathcal{S}_t} - \mathbf{\Theta}^{\mathcal{S}_t}\|_{max} \leq b_{32}\frac{|\mathcal{S}_t|^{2/m}}{n^{(m+4-\tau)/(2m)}},$$

provided that $\lambda_{n,t} \propto |\mathcal{S}_t|^{2/m} n^{-(m+4-\tau)/(2m)}$. Following a similar treatment as the proof of Theorem 3, there must exist some positive constants b_{34}, b_{35} and b_{36} such that

$$P\Big(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t\Big|\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\Big) \leq b_{34}n^{4-\tau} + b_{35}|\mathcal{S}_t|^2 n^{1-2m\phi}, \tag{28}$$

$$P\Big(\{\widehat{\text{pa}}_l \neq \text{pa}_l : l \in \widehat{\mathcal{A}}_t\}\Big|\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_t = \mathcal{A}_t\Big) \leq b_{36}|\mathcal{S}_t|^{-2}n^{4-\tau}. \tag{29}$$

Moreover, we notice that

$$\begin{aligned} P\Big(\widehat{\mathcal{E}} \neq \mathcal{E}|\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{T-1} = \mathcal{A}_{T-1}\Big) &\leq \sum_{t=0}^{T-2} P\Big(\{\widehat{\text{pa}}_l \neq \text{pa}_l : l \in \widehat{\mathcal{A}}_t\}\Big|\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_t = \mathcal{A}_t\Big) \\ &\leq b_{31}p^{2-\tau} + b_{36}(T-2)|\mathcal{S}_t|^{-2}n^{4-\tau}, \end{aligned} \tag{30}$$

where the last inequality follows from (27) and (29).

Finally, combing (26), (28) and (30), we have

$$\begin{aligned} &P(\widehat{\mathcal{G}} \neq \mathcal{G}) \\ &\leq b_{29}p^{4-\tau} + b_{30}p^2 n^{1-2m\phi} + (T-1)\Big(b_{34}n^{4-\tau} + b_{35}|\mathcal{S}_t|^2 n^{1-2m\phi}\Big) \\ &\qquad + b_{31}p^{2-\tau} + b_{36}(T-2)|\mathcal{S}_t|^{-2}n^{4-\tau} \\ &\leq b_{29}p^{4-\tau} + b_{31}p^{2-\tau} + b_{37}Tn^{4-\tau} + b_{38}Tp^2 n^{1-2m\phi}, \end{aligned}$$

where $b_{37} = 2\max\{b_{34}, b_{36}\}$ and $b_{38} = \max\{b_{30}, b_{35}\}$. Thus, when $n = \Omega\Big(T^{\frac{1}{\min\{\tau-4, 2m\phi-1\}}}(d^6 + \beta_{min}^{-2})p^{\max\{\frac{4}{m}, \frac{2}{2m\phi-1}\}}(\max\{p,n\})^{\frac{\tau-4}{m}}\Big)$, it is clear that $P(\widehat{\mathcal{G}} \neq \mathcal{G}) \to 0$ as $n \to \infty$. This completes the proof. ■

References

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, **106**:594–607, 2011.

W. Chen, M. Drton, and Y. Wang. On causal discovery with an equal-variance assumption. *Biometrika*, **106**:973–980, 2019.

D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, **3**:507–554, 2002.

G. Darmois. Analyse generale des liaisons stochastiques. *Review of the International Statistical Institute*, **21**:2–8, 1953.

D. Entner and P. Hoyer. Discovering unconfounded causal relationships using linear non-gaussian models. In *JSAI International Symposium on Artificial Intelligence*, pages 181–195. Springer, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, **9**:432–441, 2008.

M. Gao, Y. Ding, and B. Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. *Advances in Neural Information Processing Systems*, **33**, 2020.

A. Ghoshal and J. Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 1466–1475. PMLR, 2018.

N. Gnecco, N. Meinshausen, J. Peters, and S. Engelke. Causal discovery in heavy-tailed models. *Annals of Statistics*, **49**:1–25, 2021.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 585–592, 2008.

M. Ha, W. Sun, and J. Xie. PenPC: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics*, **114**:146–155, 2016.

P. Hoyer, S. Shimizu, A. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, **49**:362–378, 2008.

X. Huo and G. Székely. Fast computing for distance covariance. *Technometrics*, **58**:435–447, 2016.

A. Hyvarinen and S. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, **14**:111–152, 2013.

W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. *International Conference on Machine Learning*, pages 1742–1751, 2017.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, **8**:613–636, 2007.

M. Kalisch, M. Mächler, D. Colombo, M. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, **47**:1–26, 2012.

C. Li, X. Shen, and W. Pan. Likelihood ratio tests for a large directed acyclic graph. *Journal of the American Statistical Association*, **115**:1304–1319, 2020.

R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **107**:1129–1139, 2012.

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

W. Newey, J. Powell, and F. Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, **67**:565–603, 1999.

W. Pan, Y. Tian, X. Wang, and H. Zhang. Ball Divergence: Nonparametric two sample test. *Annals of Statistics*, **46**:1109–1137, 2018.

G. Park. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, **21**:1–34, 2020.

G. Park, S. Moon, S. Park, and J. Jeon. Learning a high-dimensional linear structural equation model via ℓ_1-regularized regression. *Journal of Machine Learning Research*, **22**:1–41, 2021.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, **101**:219–228, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference -Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017.

P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, **5**:935–980, 2011.

K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**:523–529, 2005.

A. Sanford and I. Moosa. A Bayesian network structure for operational risk modelling in structured finance operations. *Journal of the Operational Research Society*, **63**:431–444, 2012.

S. Shimizu, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, **7**:2003–2030, 2006.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. Hoyer, and K. Bollen. Directlingam: a direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, **12**:1225–1248, 2011.

W. Skitovitch. On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, **89**:217–219, 1953.

D. Spielman and S. Teng. Smoothed analysis of termination of linear programming algorithms. *Mathematical Programming*, **97**:375–404, 2003.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press, 2000.

G. Székely and M. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, **3**:1236–1265, 2009.

G. Székely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, **35**:2769–2794, 2007.

T. Tashiro, S. Shimizu, A. Hyvärinen, and T. Washio. Parcelingam: A causal ordering method robust against latent confounders. *Neural computation*, **26**:57–83, 2014.

I. Tsamardinos, L. Brown, and C. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, **65**:31–78, 2006.

C. Uhler, G. Raskitti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*, **41**:436–463, 2013.

Y. Wang and M. Drton. High-dimensional causal discovery under non-Gaussianity. *Biometrika*, **107**:41–59, 2020.

Y. Yuan, X. Shen, W. Pan, and Z. Wang. Constrained likelihood for reconstructing a directed acyclic Gaussian graph. *Biometrika*, **106**:109–125, 2019.

P. Zhang and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**:2541–2563, 2006.

X. Zheng, B. Aragam, P. Ravikumar, and E. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *In Advances in Neural Information Processing Systems (NIPS)*, 2018.

W. Zhou, X. He, W. Zhong, and J. Wang. Efficient learning of quadratic variance function directed acyclic graphs via topological layers. *Journal of Computational and Graphical Statistics*, in press, 2022.