Optimal estimation of linear non-Gaussian structural equation models

Sunmin Oh[1] **Seungsu Han**[1] **Gunwoong Park**[1,2,3]

[1]Department of Statistics, Seoul National University, Republic of Korea
[2]Institute for Data Innovation in Science, Seoul National University, Republic of Korea
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Republic of Korea

Abstract

Much of science involves discovering and modeling causal relationships in nature. Significant progress has been made in developing statistical methods for representing and identifying causal knowledge from data using Linear Non-Gaussian Acyclic Models (LiNGAMs). Despite successes in learning LiNGAMs across various sample settings, the optimal sample complexity for high-dimensional LiNGAMs remains unexplored. This study establishes the optimal sample complexity for learning the structure of LiNGAMs under a sub-Gaussianity assumption. Specifically, it introduces a structure recovery algorithm using distance covariance that achieves the optimal sample complexity, $n = \Theta(d_{in} \log \frac{p}{d_{in}})$, without assuming faithfulness or a known indegree. The theoretical findings and superiority of the proposed algorithm compared to existing algorithms are validated through numerical experiments and real data analysis.

1 Introduction

As highlighted by Eberhardt (2017); Glymour et al. (2019), under the causal Markov condition and the assumption of causal sufficiency, a directed edge in a directed acyclic graph (DAG) can be interpreted as a causal relationship. One of the established approaches to causal discovery combines directed acyclic graphs (DAGs) with linear structural equation models (SEMs) to describe the functional dependencies of effects on their causes (Pearl et al., 2000). The identifiability of linear SEMs depends on assumptions about the distribution of the additive error variables. For instance, Peters and Bühlmann (2014); Ghoshal and Honorio (2018); Park (2020) demonstrate that distribution-free linear SEMs can be identifiable under restrictions on error variances. Additionally, Shimizu et al. (2006); Zhang and Hyvärinen (2009) show that linear non-Gaussian acyclic models (LiNGAMs) can be identifiable using asymmetric conditional independence relationships.

Both identifiable linear SEMs have attracted significant attention in recent years. Specifically, for distribution-free linear SEMs, Peters and Bühlmann (2014) propose a penalized likelihood-based algorithm under equal error variance assumptions. More recently, Park (2023) introduces a computationally efficient algorithm based on topological layer-wise learning under the assumption of Gaussianity. Additionally, Loh and Bühlmann (2014); Ghoshal and Honorio (2018); Chen et al. (2019); Park et al. (2021) focus on high-dimensional ($p > n$) settings, utilizing graphical Lasso, CLIME, best-subset selection, and Lasso. Finally, Gao et al. (2022) present a sample optimal algorithm under the assumptions of Gaussianity and equal error variances, while suggesting that both assumptions can be relaxed.

Several algorithms for learning LiNGAMs have also been developed. Notably, Shimizu et al. (2006) propose an iterative search algorithm that recovers the ordering of a LiNGAM using linear independent component analysis and permutation. This is improved by Shimizu et al. (2011), which introduces a more computationally efficient algorithm that leverages pairwise statistics to learn linear non-Gaussian DAGs. Hyvärinen and Smith (2013) further extend this work, by iteratively identifying pairwise causal ordering using likelihood ratio tests. For high-dimensional settings, Wang and Drton (2020); Zhao et al. (2022) develop consistent algorithms with upper bounds on sample com-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

plexity. Additional extensions of LiNGAMs have been proposed for time-series data (Gong et al., 2017), the case of polytrees (Tramontano et al., 2022), nonlinear models (Hoyer et al., 2009; Zhang and Hyvärinen, 2009), and models with latent variables (Shimizu and Hyvärinen, 2007; Maeda and Shimizu, 2020).

Despite the success of learning LiNGAMs across various sample settings, a tight characterization of the optimal sample complexity for learning LiNGAMs from observational data remains unexplored. This contrasts with distribution-free linear SEMs, where optimal sample complexity has been established under the Gaussianity assumption. Notably, the existing upper bounds on high-dimensional sample complexity proposed by Wang and Drton (2020); Zhao et al. (2022) are far from the information-theoretic lower bounds for (Gaussian) linear SEMs, as detailed in Gao et al. (2022); Ghoshal and Honorio (2017). Hence, its application could be limited in fields like biology, particularly in gene expression data, which faces the challenge of extremely small sample sizes when applying high-dimensional causal discovery. This raises a fundamental question: Is it possible to establish the optimal sample complexity for learning the structure of LiNGAMs?

This study demonstrates that an optimal learning is achievable for sub-Gaussian LiNGAMs, where error distributions are non-Gaussian but sub-Gaussian. Specifically, we propose an optimal LiNGAM learning algorithm based on the distance covariance and best-subset-selection method, with the optimal sample complexity of $n = \Theta(d_{in} \log \frac{p}{d_{in}})$. This involves a novel analysis of the LiNGAM learning algorithm, sharpening the existing upper bounds on sample complexity, deriving a lower bound, and proving their alignment. Here, n is the number of samples, p is the number of nodes, and d_{in} is the maximum indegree of a graph. To the best of our knowledge, this is the first result that establishes the optimal sample complexity for learning LiNGAMs in high-dimensional regimes, with the upper bound being optimal up to constant factors.

The rest of this paper is organized as follows. Section 2 introduces the notation and problem settings, including a detailed explanation of LiNGAMs. It also provides a comparison of existing approaches for learning linear SEMs, focusing on their required assumptions and theoretical properties. Section 3 presents Proposition 1 and Theorem 2, which ensure that LiNGAMs can be recovered by performing independence tests. Section 4 proposes an optimal algorithm for high-dimensional sub-Gaussian LiNGAMs, utilizing the distance covariance-based independence test, and presents the theoretical guarantees for the proposed algorithm. Sections 5 and 6 assess the performance of the proposed and state-of-the-art algorithms using synthetic and real sociological survey data. Finally, Section 7 discusses potential avenues for future research.

2 Preliminaries

This section introduces the problem settings and some necessary notations. Subsequently, it discusses related works for learning linear SEMs.

2.1 Problem Settings

Let $G = (V, E)$ be a DAG with a set of nodes $V = \{1, 2, ..., p\}$ and a set of directed edges $E \subset V \times V$. A directed edge from node j to k is denoted by (j, k) or $j \to k$. The *parent* set of node k, denoted by $\mathrm{Pa}(k)$, consists of all nodes j such that $(j, k) \in E$. The *child* set, denoted by $\mathrm{Ch}(j)$, consists of all nodes k such that $(j, k) \in E$. The sets $\mathrm{An}(k)$ and $\mathrm{Nd}(k)$ denote the *ancestors* and *non-descendants* of node k, respectively.

There exists an *ordering* $\pi = (\pi_1, \pi_2, ..., \pi_p)$ of a graph that indicates directions of edges such that j appears before k in the ordering for every directed edge $(j, k) \in E$. There also exists the set of *topological layers* $\cup_{t=0}^{T-1} \mathcal{A}_t = V$, where the layer $\mathcal{A}_t$ consists of nodes whose longest distance to a source node is t. Each node is assigned to only one layer, and for each $j \in \mathcal{A}_r$, its parents $\mathrm{Pa}(j)$ are contained in $S_{r-1} = \cup_{t=0}^{r-1} \mathcal{A}_t$. Hence, learning the topological layers leads to inferring the ordering, and hence, the directions of the edges.

Let $X := (X_j)_{j \in V}$ be a set of random variables with a probability distribution over the nodes in G. For any subset S of V, let $X_S := \{X_j : j \in S \subset V\}$. Additionally, for any node $j \in V$, $P(X_j \mid X_S)$ denotes the conditional distribution of a variable X_j given a random vector X_S. Then, the joint distribution of X is defined as $P(X) = \prod_{j \in V} P(X_j \mid X_{\mathrm{Pa}(j)})$.

In this study, we assume an independent and identically distributed sample matrix $\mathbf{x}^{1:n} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$ of size $n \times p$ from a given model. The notation $\widehat{\cdot}$ denotes an estimate based on the sample $\mathbf{x}^{1:n}$. Finally, the detailed proofs for our theoretical results are provided in Supplementary.

2.2 Sparse Sub-Gaussian LiNGAMs

A linear non-Gaussian SEM (LiNGAM) is constructed by the following linear structural equations: For any node $j \in V$,

$$X_j = \sum_{k \in \mathrm{Pa}(j)} \beta_{k,j} X_k + \epsilon_j, \tag{1}$$

where $\epsilon = (\epsilon_j)_{j \in V}$ are independent continuous non-Gaussian errors. Additionally, $\beta_{k,j} \in \mathbb{R}$ is the linear weight of an edge from X_k to X_j in which $\beta_{k,j} = 0$ if and only if $(k, j) \notin E$.

Shimizu et al. (2006) show that LiNGAMs are identifiable by reformulating (1) as a form of non-Gaussian linear independent component analysis model, which is known to be identifiable as in Comon (1994); Eriksson and Koivunen (2004). Shimizu et al. (2011) apply Darmois-Skitovitch theorem (Darmois, 1953; Skitovitch, 1953) to reconstruct the ordering of a LiNGAM. Specifically, an exogenous variable, whose existence is ensured by the acyclicity assumption, is identified using Darmois-Skitovitch theorem. Subsequently, the effects of the exogenous variables are removed from other variables through least squares regression, revealing another exogenous variable. By iterating this process, the complete ordering can be recovered.

This study focuses on *sub-Gaussian LiNGAMs* to achieve the optimal sample complexity for structure learning. Specifically, the models allow continuous, sub-Gaussian, but non-Gaussian error distributions. For instance, if all error distributions are uniform, then it is a sub-Gaussian LiNGAM. Additionally, we consider *sparse* models with a small maximum indegree, rather than a small maximum degree of the moralized graph, as explored in Ghoshal and Honorio (2018); Zhao et al. (2022).

2.3 Relevant Algorithms for Linear SEMs

A significant open question in the literature on DAG recovery is the optimal sample complexity of learning a linear SEM. The derivation of upper bounds of the sample complexity for this problem has been significant progress (e.g., Loh and Bühlmann, 2014; Ghoshal and Honorio, 2018; Chen et al., 2019; Park et al., 2021; Wang and Drton, 2020; Zhao et al., 2022; Hwang et al., 2023). Additionally, the lower bounds of the sample complexity for various DAG models have been investigated (e.g., Ghoshal and Honorio, 2017; Gao et al., 2022).

Consequently, for Gaussian linear SEMs, Gao et al. (2022) establish the high-dimensional consistency with the optimal sample complexity $n = \Theta(d_{in} \log \frac{p}{d_{in}})$ under the equal error variance and known maximum indegree assumptions, where $f(x) = \Theta(g(x))$ if and only if $f(x) = O(g(x))$ and $g(x) = O(f(x))$. However, for LiNGAMs, an optimal learning method has not yet been developed.

Although not optimal, several attempts have been made to establish high-dimensional consistency for LiNGAMs. For example, Zhao et al. (2022) develop an algorithm with sample complexity $n = \Omega(T^{c_1} d^6 \log(p)^{c_2})$, for some positive constants $c_1 > 1$ and $c_2 > 3$, where T is the total number of topological layers and d is the maximum degree of the moralized graph. Similarly, Wang and Drton (2020) establish high-dimensional consistency with sample complexity $n = \Omega((\log p)^{2K})$ for some positive integer $K > 1$ under small indegree conditions. However, these sample complexities are far from the lower bound of the sample complexity for sub-Gaussian LiNGAMs, $n = O(d_{in} \log \frac{p}{d_{in}})$ (see details in Section 4.1.3).

Furthermore, these methods rely on certain restrictive assumptions. Specifically, Wang and Drton (2020) require the parental faithful assumption, which is less strict than the faithfulness assumption but still restrictive. Zhao et al. (2022) impose the incoherence assumption which is necessary for accurately recovering the precision matrix using graphical Lasso (Friedman et al., 2008; Ravikumar et al., 2011).

3 New Properties of LiNGAMs

This section introduces novel conditional independence properties of LiNGAMs, which enables us to directly determine the causal directions in LiNGAMs. Specifically, we show that (i) a variable becomes exogenous when its projection onto a non-descendant set is subtracted, if and only if the set contains all of its parents. Hence, whether a non-descendant set contains the parent set, $\mathrm{Pa}(j)$, can be inferred by determining whether the variable X_j can be exogenous after subtracting the projection onto a subset of the non-descendant set. Consequently, the element of the ordering that directly follows the non-descendant set can be identified. Additionally, we show that (ii) the minimal subset of the non-descendant set that makes a variable exogenous corresponds to its parent set.

Precisely, let $\Sigma = \mathbb{E}(XX^T)$ be the covariance matrix of the centered random vector X and $\widehat{\Sigma}$ be its sample covariance matrix. Let $\widehat{\Sigma}_{A,B}$ be the $|A| \times |B|$ submatrix of $\widehat{\Sigma}$ corresponding to random vectors X_A and X_B, where $A, B \subseteq V$. For $j \in V \setminus C$ and $C \subset V$, let $\Sigma_{C,j}$ be the subvector composed of the entries in places (c, j) for $c \in C$. Denote the residual of X_j, where X_j is regressed onto X_C, as $e_{j,C} = X_j - \Sigma_{j,C}(\Sigma_{C,C})^{-1} X_C$. When $C = \emptyset$, let $e_{j,C} = X_j$. Finally, for $j, k \in V$ and $C \subset V \setminus \{j, k\}$,

$$r_C(j, k) = e_{j,C} - \frac{\mathrm{Cov}(e_{j,C}, e_{k,C})}{\mathrm{Var}(e_{k,C})} e_{k,C}.$$

Proposition 1. *Let $P(X)$ be generated from a LiNGAM (1) with DAG G and the topological layers $\cup_{t=0}^{T-1} \mathcal{A}_t$. For any $r \in \{1, 2, ..., T-1\}$, $j \in \mathcal{A}_r$, $k \in V \setminus \cup_{t=0}^{r} \mathcal{A}_t$, and $S_{r-1} = \cup_{t=0}^{r-1} \mathcal{A}_t$,*

(i) *there exists* $C \subset S_{r-1}$ *satisfying* $Pa(j) \subset C$, *and then* $e_{j,C} \perp\!\!\!\perp r_C(\ell, j)$ *for all* $\ell \in V \setminus (S_{r-1} \cup \{j\})$.

(ii) *if* $C \subset S_{r-1}$ *satisfies* $e_{j,C} \perp\!\!\!\perp X_\ell$ *for all* $\ell \in S_{r-1}$, *then* $Pa(j) \subset C$.

(iii) *for any* $C \subset S_{r-1}$, *there exists* $\ell \in V \setminus (S_{r-1} \cup \{k\})$ *such that* $e_{k,C} \not\perp\!\!\!\perp r_C(\ell, k)$.

Proposition 1 ensures that the set of variables required to make a variable exogenous is reduced from the entire set of preceding elements in the causal ordering to only its parent set. This reduction provides a distinct advantage for LiNGAM learning compared to the approaches in Shimizu et al. (2011) and Zhao et al. (2022) (see details in Lemma 1 of Shimizu et al., 2011 and Theorem 1 of Zhao et al., 2022). Additionally, Proposition 1 does not rely on (parental) faithfulness, distinguishing it from the approach in Wang and Drton (2020) and providing another advantage for LiNGAM learning (see details in Theorem 1 of Wang and Drton, 2020).

Specifically, all topological layers of a LiNGAM can be reconstructed by iteratively applying Proposition 1 (i) and (iii) in a top-down manner. By combining these propositions, we can distinguish between the earlier and later orderings. Additionally, Proposition 1 (ii) reveals that the minimal conditioning set $C \subset S_{r-1}$, for which $e_{j,C}$ is independent of X_m for all $m \in S_{r-1}$, is $\mathrm{Pa}(j)$. Then, integrating all parts of Proposition 1, enables the simultaneous recover of both the topological layers and the parents of each node by identifying the minimal conditioning set C, where $e_{j,C}$ is independent of both $r_C(\ell, j)$ and X_m for all $\ell \in V \setminus (S_{r-1} \cup \{j\})$ and $m \in S_{r-1}$. This insight motivates the use of dependency relationships between (residualized) variables and conditioning sets for the following property, which helps recover the true graph structure.

Theorem 2. *Let* $P(X)$ *be generated from a LiNGAM* (1) *with DAG* G *and the topological layers* $\cup_{t=0}^{T-1} \mathcal{A}_t = V$. *Consider any* $r \in \{1, 2, ..., T-1\}$ *and* $S_{r-1} = \cup_{t=0}^{r-1} \mathcal{A}_t$. *Then*

$$\begin{aligned} &\mathcal{A}_0 = \left\{ j \in V : X_j \perp\!\!\!\perp e_{k,\{j\}} \text{ for all } k \in V \setminus \{j\} \right\}, \text{ and} \\ &\mathcal{A}_r = \{ j \in V \setminus S_{r-1} : \exists C_j \subset S_{r-1} \text{ s.t.} \\ &\quad X_k \perp\!\!\!\perp e_{j,C_j} \text{ for all } k \in S_{r-1} \text{ and} \\ &\quad e_{j,C_j} \perp\!\!\!\perp r_{C_j}(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\}) \}. \end{aligned} \tag{2}$$

Moreover, for each $j \in \mathcal{A}_r$ *and the corresponding set*

$$\begin{aligned} C_j = \{ &C \subset S_{r-1} : X_k \perp\!\!\!\perp e_{j,C} \text{ for all } k \in S_{r-1} \\ &\text{and } e_{j,C} \perp\!\!\!\perp r_C(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\}) \}, \end{aligned}$$

$Pa(j) \subset C \subset Nd(j)$ *for any* $C \in C_j$.

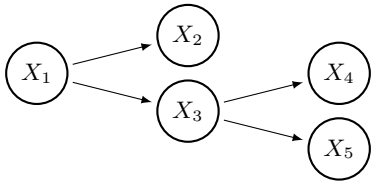


Figure 1: Tree LiNGAM

Theorem 2 ensures that the true ordering can be correctly identified, by substituting $\pi_r \in \pi$ as a true ordering and setting $S_{r-1} = \{\pi_1, ..., \pi_{r-1}\}$ with $S_0 = \emptyset$. Additionally, for each selected node j, the minimal set $C \in C_j$ corresponds to $\mathrm{Pa}(j)$. Consequently, the graph can be recovered more efficiently based on this new independence property.

To provide an intuition, we explain how Theorem 2 can be applied to recover the graph using a 5-node tree LiNGAM in Figure 1. First, $\mathcal{A}_0 = \{1\}$ can be determined because X_1 is the only exogenous variable. Next, the other variables are projected onto X_1. Then, $e_{2,\{1\}} = \epsilon_2$ and $e_{3,\{1\}} = \epsilon_3$ are exogenous, whereas $e_{4,\{1\}} = \beta_{3,4}\epsilon_3 + \epsilon_4$ and $e_{5,\{1\}} = \beta_{3,5}\epsilon_3 + \epsilon_5$ are not, because they are dependent of

$$r_{\{1\}}(3,4) = \epsilon_3 - \frac{\beta_{3,4}\mathrm{Var}(\epsilon_3)}{\beta_{3,4}^2\mathrm{Var}(\epsilon_3) + \mathrm{Var}(\epsilon_4)}(\beta_{3,4}\epsilon_3 + \epsilon_4), \text{ and}$$

$$r_{\{1\}}(3,5) = \epsilon_3 - \frac{\beta_{3,5}\mathrm{Var}(\epsilon_3)}{\beta_{3,5}^2\mathrm{Var}(\epsilon_3) + \mathrm{Var}(\epsilon_5)}(\beta_{3,5}\epsilon_3 + \epsilon_5),$$

respectively. Hence, $\mathcal{A}_1 = \{2, 3\}$ and $\mathrm{Pa}(2) = \mathrm{Pa}(3) = \{1\}$. Subsequently, since $e_{4,\{3\}} = e_{4,\{2,3\}} = e_{4,\{1,3\}} = e_{4,\{1,2,3\}} = \epsilon_4$ is exogenous, we have $4 \in \mathcal{A}_2$ and the minimal set $\{3\}$ is identified as $\mathrm{Pa}(4)$. Similarly, $5 \in \mathcal{A}_2$ and $\mathrm{Pa}(5) = \{3\}$, which completes the graph recovery.

It is important to note that, to learn the complete graph structure, projections onto at most two predictors are sufficient. Specifically, $r_1(3,4)$ is the residual of $e_{3,1}$ when projected onto $e_{4,1}$, which is equivalent to the residual of X_3 when projected onto the subspace spanned by X_1 and X_4. This reduction is enabled by Theorem 2, which reduces the number of variables necessary to ensure the exogeneity of a variable. In comparison to Shimizu et al. (2011) or Zhao et al. (2022), the number of predictors required for projection decreases from $p-1$ or d to $d_{in}+1$.

4 Algorithm

This section introduces a consistent algorithm for LiNGAMs even in high-dimensional sparse settings, applying the characteristics of LiNGAMs specified in Theorem 2. It is important to note that the use of the new independence property allows the proposed algorithm to achieve sample optimality under certain

Algorithm 1: Optimal LiNGAM Algorithm

Input : n i.i.d. samples $\mathbf{x}^{1:n}$ and significance level α

Output: Estimated graph, $\widehat{G} = (V, \widehat{E})$

Step 1) Source nodes estimation

$$\widehat{\mathcal{A}}_0 = \left\{ j \in V : \mathbf{x}_j \perp\!\!\!\perp_{\text{test}} \hat{e}_{k,\{j\}} \text{ for all } k \in V \setminus \{j\} \right\}.$$

Step 2) Directed edges (parent) estimation

Initialize $\mathcal{R} = \widehat{\mathcal{A}}_0$.

while $V \setminus \mathcal{R} \neq \emptyset$ **do**
  for $q \in \{1, 2, ..., |\mathcal{R}|\}$ **do**
    Set $\mathcal{R}_0 = \emptyset$
    for $j \in V \setminus \mathcal{R}$ **do**
      if $\min_{|C|=q, C \subset \mathcal{R}} \widehat{\mathcal{S}}(j, C) = 0$ *using* (3) **then**
        Update $\mathcal{R}_0 = \mathcal{R}_0 \cup \{j\}$ and $\widehat{\text{Pa}}(j) = \arg\min_{|C|=q, C \subset \mathcal{R}} \widehat{\mathcal{S}}(j, C)$
      end
    end
    if $\mathcal{R}_0 \neq \emptyset$ **then**
      Update $\mathcal{R} = \mathcal{R} \cup \mathcal{R}_0$
      Break
    end
  end
end

Return: $\widehat{E} = \{(k, j) : j \in V, k \in \widehat{\text{Pa}}(j)\}$

conditions, such as sub-Gaussianity, without assuming a known d_{in}. Additionally, it does not require common assumptions for high-dimensional LiNGAMs, such as the incoherence and the parental faithfulness assumptions.

Specifically, Step 1) of Algorithm 1 estimates the first topological layer using the idea of Lemma 1 of (Shimizu et al., 2011). The proposed algorithm identifies exogenous variables by using non-Gaussianity and independence relationship.

$$\widehat{\mathcal{A}}_0 = \left\{ j \in V : \mathbf{x}_j \perp\!\!\!\perp_{\text{test}} \hat{e}_{k,\{j\}} \text{ for all } k \in V \setminus \{j\} \right\},$$

where $\perp\!\!\!\perp_{\text{test}}$ indicates that independence is tested and $\hat{e}_{k,\{j\}}$ follows the notation introduced in Equation (4). Any valid independence test can be applied, such as Hilbert-Schmidt independence criterion (Gretton et al., 2007) and distance covariance measure (Székely et al., 2007; Székely and Rizzo, 2013).

Step 2) estimates the remaining topological layers and their parents using the independence properties in Theorem 2 and the following dependency score. The score is defined as the sum of two components: the number of preceding elements in the ordering tested for dependence with $\hat{e}_{j,C}$, and the number of residualized variables also tested for dependence with $\hat{e}_{j,C}$. Precisely, for $j \in V$ and a small $C \subset V \setminus \{j\}$,

$$\begin{aligned} \widehat{\mathcal{S}}(j, C) := &\#\{k \in \mathcal{R} : \hat{e}_{j,C} \not\perp\!\!\!\perp_{\text{test}} \mathbf{x}_k\} \\ &+ \#\{\ell \in V \setminus (\mathcal{R} \cup \{j\}) : \hat{e}_{j,C} \not\perp\!\!\!\perp_{\text{test}} \hat{r}_C(\ell, j)\}, \end{aligned} \tag{3}$$

where

$$\begin{aligned} \hat{e}_{j,C} &= \mathbf{x}_j - \mathbf{x}_C (\widehat{\Sigma}_{C,C})^{-1} \widehat{\Sigma}_{Cj}, \\ \hat{r}_C(j, k) &= \hat{e}_{j,C} - \frac{\widehat{\Sigma}_{j,k} - \widehat{\Sigma}_{j,C} (\widehat{\Sigma}_{C,C})^{-1} \widehat{\Sigma}_{C,k}}{\widehat{\Sigma}_{k,k} - \widehat{\Sigma}_{k,C} (\widehat{\Sigma}_{C,C})^{-1} \widehat{\Sigma}_{C,k}} \hat{e}_{k,C}. \end{aligned} \tag{4}$$

This score is constructed upon the independence property in Theorem 2. Specifically, assuming independence tests are perfectly accurate, given that the partial ordering is correctly recovered in prior iterations, there exists a subset $C \subset \mathcal{R}$ for which $\widehat{\mathcal{S}}(j, C) = 0$ if and only if j is a source node in the subgraph obtained after removing the elements of the partial ordering recovered in prior iterations. Furthermore, $\widehat{\mathcal{S}}(j, C) = 0$ only if $\text{Pa}(j) \subset C \subset \mathcal{R}$. Hence, the proposed algorithm accurately determines the directed edges by checking which set of nodes makes the score zero.

To reduce the computational cost of finding the minimal node set, the algorithm incrementally expands the conditioning set during its search. It then selects source nodes in the subgraph whose indegree in the entire graph is the smallest among the remaining source nodes at each iteration. As a result, the estimated conditioning set C is guaranteed to be minimal, satisfying $\widehat{\mathcal{S}}(j, C) = 0$, and hence $C = \text{Pa}(j)$. Furthermore, the size of the conditioning sets, denoted by q, should be less than or equal to the maximum indegree d_{in}. This process is repeated until the remaining nodes are empty, i.e., $V \setminus \mathcal{R} = \emptyset$.

4.1 Theoretical Guarantees

This section demonstrates the sample optimality of Algorithm 1 for learning sub-Gaussian LiNGAMs when the independence test is performed using distance covariance, as proposed in Székely et al. (2007); Székely and Rizzo (2013); Zhao et al. (2022). It outlines the required assumptions and sample size for Algorithm 1 and presents the minimum sample size needed for any estimator to recover the graph.

4.1.1 Required Assumptions

We begin by discussing the assumption imposed on the covariance matrix, $\Sigma \succ 0$.

Assumption 3 (Dependency Assumption)**.** *There exists a positive constant $\lambda > 0$ such that*

$$\lambda^{-1} \leq \Lambda_{\min}(\Sigma) \leq \Lambda_{\max}(\Sigma) \leq \lambda,$$

where $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ are the minimum and maximum eigenvalues of a matrix A, respectively.

Assumption 3 ensures that the variables maintain balanced dependence and marginal variances. It prevents situations where variables become nearly linearly dependent (with eigenvalues approaching zero) or overly dominant (resulting in very large eigenvalues). This assumption is common in high-dimensional settings to preserve well-behaved covariance structures, ensuring numerical stability and robustness in performance of the algorithm. Consequently, it frequently appears in the relevant literature, such as Ghoshal and Honorio (2018); Chen et al. (2019); Park (2020); Gao et al. (2022), where high-dimensional (sub-)Gaussian linear SEMs are studied.

Assumption 4 (Minimal Signal Assumption)**.** *There exists a positive constant τ_1 defined as follows:*

$$\tau_1 = \min\{T(e_{j,C}, r_C(\ell, j)) \neq 0 : j, \ell \in V, C \subset Nd(j) \setminus \{\ell\}\} \wedge \min\{T(e_{j,C}, X_k) \neq 0 : j, k \in V, C \subset V \setminus \{j\}\},$$

where T is the test statistic based on distance covariance, whose detailed definition is in Section 11.1 of Supplementary.

Assumption 4 ensures that the dependence between variables is sufficiently strong to be detected, thereby improving the accuracy of the independence test. It mirrors Condition (C2) from Li et al. (2012), which is crucial for feature screening using distance covariance.

4.1.2 Consistency

Armed with Assumptions 3 and 4, we can establish a main result on the optimal sample complexity for the proposed algorithm.

Theorem 5. *Consider a sub-Gaussian LiNGAM (1) with $d_{in} \leq \frac{p}{2}$. Suppose that Assumptions 3 and 4 are satisfied. Let $\widehat{G} = (V, \widehat{E})$ be the estimated graph from Algorithm 1 with the significance level of the independence test $\alpha_n = 2(1 - \Phi(\sqrt{n}\epsilon))$ for any $\epsilon_0 \in (0, \tau_1/2)$. Then, for some positive constants D_1 and D_2,*

$$\Pr(\widehat{G} = G) \geq 1 - D_1 \left(\frac{p}{d_{in}}\right)^{d_{in}} exp\left(-\frac{D_2 \epsilon_0^4}{\lambda^{36}} n\right).$$

Theorem 5 claims that if the sample size scales to $n = \Omega(d_{in} \log \frac{p}{d_{in}})$, the proposed algorithm accurately learns a sub-Gaussian LiNGAM with high probability under appropriate conditions. A key idea of the proof is to show a consistent test statistic estimator with an error bound that exponentially converges to zero. Additionally, it addresses the high-dimensional problem by incrementally searching over the conditioning set, ensuring that its size is bounded by d_{in}. This approach allows the sample complexity to scale with d_{in} without prior knowledge of d_{in}, whereas the sample complexities of comparison algorithms depend on d or require d_{in} to be known (Zhao et al., 2022; Gao et al., 2022).

It is important to note that the significance level α_n converges to zero as n increases to ensure that the type I error approaches zero. Additionally, it is evident that as τ_1 increases, a larger ϵ_0 can be utilized, resulting in faster convergence of the estimated graph to the true graph. This is logically consistent, as a larger minimum distance covariance between dependent (residualized) variables enhances the clarity of the functional relationships, which in turn eases graph recovery.

4.1.3 Sample Optimality

It is now focused on the sample optimality by showing the minimum sample size for successful sub-Gaussian LiNGAM learning required by *any* estimator. Let $\mathcal{M}_{p,d_{in}}(\lambda)$ be the class of p-node sub-Gaussian LiNGAMs where the maximum indegree is $d_{in} \leq \frac{p}{2}$, and Assumption 3 holds with a positive constant λ.

Lemma 6. *Suppose that $G(M)$ denotes a true DAG corresponding to a model $M \in \mathcal{M}_{p,d_{in}}(\lambda)$. For any positive constant $\delta \in (0, \frac{1}{2})$, there are some positive constants K_1 and K_2 such that if the sample size is*

$$n \leq (1 - 2\delta) K_1 \frac{d_{in} \log(p/d_{in})}{\log(\lambda)},$$

then, for any estimator $\widehat{G}$,

$$\sup_{M \in \mathcal{M}_{p,d_{in}}(\lambda)} \Pr(\widehat{G} \neq G(M)) \geq \delta - \frac{K_2}{p d_{in} \log(p/d_{in})}.$$

Lemma 6 asserts that if $n \leq c d_{in} \log \frac{p}{d_{in}}$ for some positive constants c, all estimators fail to recover a graph with high probability (i.e., the minimum sample size is $n = \Omega(d_{in} \log \frac{p}{d_{in}})$). A key idea of the proof is analogous to the previous works in Ghoshal and Honorio (2017); Gao et al. (2022), which analyze linear SEMs with restrictions on the error distribution or graph structures. However, our result considers sub-Gaussian distributions, excluding Gaussian, and imposes no restrictions on error variances.

In summary, Theorem 5 combined with Lemma 6 demonstrates that the upper and lower bounds of the sample complexity align up to numerical constant factors. Hence, the proposed algorithm is sample optimal concerning its dependence on p and d_{in}.

Corollary 7. *Algorithm 1 achieves the optimal sample complexity $n = \Theta\left(d_{in} \log \frac{p}{d_{in}}\right)$ for sub-Gaussian LiNGAMs under appropriate conditions.*

5 Numerical Experiments

This section evaluates the numerical performance of Algorithm 1 (OptLiNGAM) and compares it with well-known LiNGAM learning methods, including ICALiNGAM (Shimizu et al., 2006), DirectLiNGAM (Shimizu et al., 2011), and TL (Zhao et al., 2022), as well as other linear SEM learning methods, such as LISTEN (Ghoshal and Honorio, 2018) and HLSM (Park et al., 2021). Specifically, OptLiNGAM utilizes the independence test based on distance covariance with the significance level $\alpha = 2(1 - \Phi(\sqrt{c_0 n}))$, as shown in Theorem 5. The regularization parameters for LISTEN, HLSM, and TL, are implemented according to the recommended settings provided in the respective papers. Similarly, the significance level of independent tests in TL is set as $\alpha = 0.01$ as recommended in Zhao et al. (2022).

Various simulation settings are considered, similar to those used by Wang and Drton (2020) and Zhao et al. (2022), where high-dimensional LiNGAM recovery was considered. Specifically, three types of hub graphs are generated with $d_{in} \in \{1, 2, 3\}$, where the number of hub nodes corresponds to d_{in}, $\lfloor \log p \rfloor$ nodes are isolated, and the remaining nodes are children of the hub nodes. Error terms are drawn from Beta$(0.5, 0.5)$, with nonzero edge weights sampled uniformly from $[-1.5, -0.5] \cup [0.5, 1.5]$ and the data generation process is repeated 30 times. Further simulation results in other settings and the corresponding details are provided in Section 14 of Supplementary.

All algorithms are evaluated using the Matthews correlation coefficient (MCC) and the normalized structural Hamming distance (HM) (Tsamardinos et al., 2006). MCC measures the accuracy of the estimated directed edges, ranging from -1 to $+1$, where $+1$ indicates a perfect prediction, 0 represents an average random prediction, and -1 signifies an inverse prediction. Additionally, HM quantifies the similarity between the estimated and the true DAG, with lower HM values indicating better estimation accuracy.

5.1 Verification of Consistency

We present the average MCC of the considered algorithms by varying sample size $n \in \{100, 200, ..., 500\}$ in Figure 2. Specifically, Figure 2 (a) shows that the proposed algorithm consistently recovers the graph. Specifically, as n increases, the average MCC converges to 1, indicating perfect graph recovery, with faster convergence observed for smaller maximum indegree values. This demonstrates that OptLiNGAM requires fewer samples to recover sparse graphs with lower maximum indegree. Therefore, it validates Theorem 5, confirming the consistency of OptLiNGAM and the sample complexity depending on d_{in}.

Figures 2 (b) - (d) compare the algorithms and show that OptLiNGAM significantly outperforms the comparison methods as sample size increases. Notably, some comparison methods seem to fail to recover the graph regardless of sample size due to restrictive required conditions. Hence, this highlights the advantages of the proposed algorithm: being sample optimal and not requiring restrictive conditions.

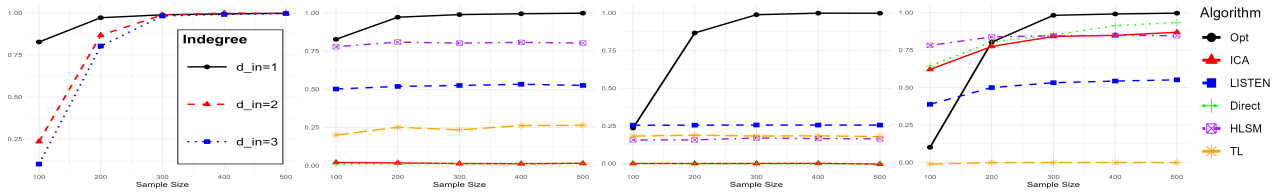
5.2 Large-Scale Graph Structure Learning

To further verify the sample efficiency of OptLiNGAM, we compare its numerical performance with the other algorithms in large-scale models. Specifically, we focus on learning sparse models with $d_{in} = 1$ for 100 and 200 nodes, while keeping the sample size fixed at 200.

Table 1: Average measures of all algorithms in learning sparse models with standard errors in parentheses. The best values are highlighted in bold.

(n, p)	Method	MCC	HM
(200,100)	Opt	**0.8959**(0.0810)	**0.0018**(0.0013)
	ICA	0.6024(0.0550)	0.0131(0.0034)
	Direct	0.5812(0.0508)	0.0157(0.0037)
	TL	-0.0004(0.0006)	0.0096(0.0001)
	LISTEN	0.3709(0.0285)	0.0083(0.0002)
	HLSM	0.7872(0.0269)	0.0057(0.0010)
(200,200)	Opt	**0.8592**(0.0993)	**0.0012**(0.0008)
	ICA	0.4686(0.1557)	0.0141(0.0094)
	Direct	0.4439(0.0385)	0.0164(0.0029)
	TL	-0.0002(0.0002)	0.0049(0.0000)
	LISTEN	0.2656(0.0261)	0.0045(0.0001)
	HLSM	0.7308(0.0266)	0.0040(0.0006)

As shown in Table 1, OptLiNGAM recovers most directed edges even in large-scale graph setting. As expected, it outperforms the comparison algorithms in both MCC and HM. Specifically, for $n = 200$ and $p = 200$, OptLiNGAM achieves the average MCC of 0.8592 and the average HM of 0.0012. In comparison, the second-best algorithm, HLSM–which is not designed for LiNGAM recovery–has the average MCC of 0.7308 and the average HM of 0.0040. Among the LiNGAM learning algorithms, ICA achieves the average MCC of 0.4686 and the average HM of 0.0141. Notably, TL predominantly estimates empty graphs, leading to an MCC close to zero and a negligibly small HM value under sparse true graph settings. Therefore, the simulation results highlight the need for improvement in existing LiNGAM learning algorithms in large-scale and high-dimensional model settings, emphasizing the advantages of the proposed algorithm.



(a) OptLiNGAM (b) $d_{in} = 1$ (c) $d_{in} = 2$ (d) $d_{in} = 3$

Figure 2: Average Matthews correlation coefficient for 50-node sparse hub graphs with $d_{in} \in \{1, 2, 3\}$.

6 Real Data Analysis

We apply the proposed algorithm to real-world sociological data from the General Social Survey (http://www.norc.org/GSS+Website), a national survey of U.S. adults. The data set comprises six variables: (1) Father's Education, (2) Father's Occupation, (3) Number of Siblings, (4) Son's Education, (5) Son's Occupation, and (6) Son's Income. Figure 3 (a) illustrates the true graph based on domain knowledge (see details in Duncan et al., 1972). Notably, the bi-directed edges between nodes (1), (2), and (3) appear due to the presence of latent confounders. Consequently, the bi-directed edges are excluded from the interpretation of the results.

This data was analyzed by DirectLiNGAM algorithm in Shimizu et al. (2011). Specifically, they focused on samples of white, male individuals with a non-farm background, aged 35-44, who were in the labor force at the time of the survey, and were surveyed for 45 years, from 1972 to 2006. In contrast, this analysis employs data for five years, from 2002 to 2006, consisting of 355 observations, to demonstrate the ability of OptLiNGAM for sample efficient LiNGAM recovery.

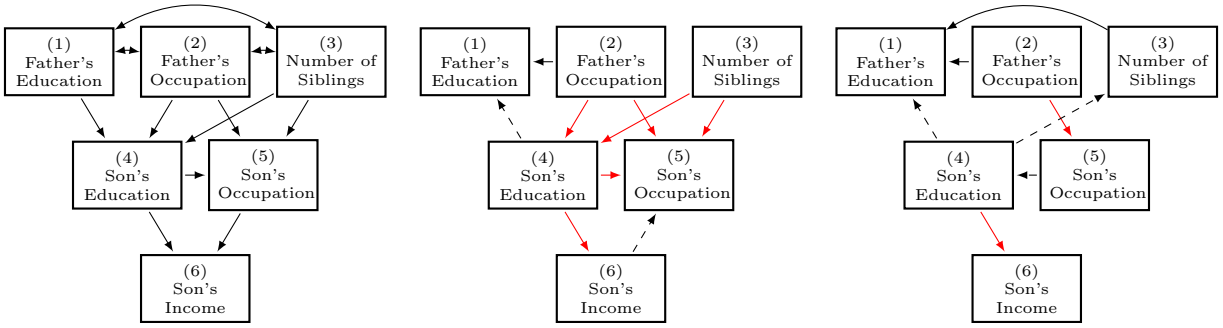
Figures 3 (b) and (c) show the graphs estimated by OptLiNGAM and DirectLiNGAM, respectively, where $\alpha = 2(1 - \Phi(\sqrt{0.01n}))$. It shows that the proposed algorithm demonstrates remarkable performance with fewer observations, thereby confirming its sample efficiency. Specifically, OptLiNGAM successfully identifies most causal relationships between variables, except for two reversed edges, (4, 1) and (6, 5). In contrast, DirectLiNGAM fails to capture many important links, such as (2, 4) and (3, 5), which OptLiNGAM correctly identifies. Additionally it falsely determines the direction of edges, including (4, 3) and (5, 4), which are accurately detected by OptLiNGAM. Consequently, OptLiNGAM shows superior performance, achieving a Hamming distance of 2, whereas DirectLiNGAM achieves a distance of 6. A detailed comparison with other LiNGAM recovery approaches is provided in Section 15 of Supplementary.

7 Conclusion

This study develops a sample optimal approach that directly recovers the parent set of each node in a top-down manner by introducing novel independence properties of LiNGAMs. Leveraging the characteristics of LiNGAMs, the set of variables required to identify a variable as exogenous is narrowed from the entire set of preceding elements to its parent set. Consequently, the proposed algorithm attains efficient sample complexity that depends on d_{in}, while avoiding common assumptions in high-dimensional LiNGAMs, such as the incoherence assumption or the parental faithfulness. Additionally, this study derives a lower bound on the sample complexity for learning LiNGAMs and confirms the sample optimality of the proposed algorithm, even without prior knowledge of d_{in}.

The proposed algorithm requires a significance level for independent testings, with an appropriate range provided in Theorem 5. A widely used practical approach for selecting the significance level is cross-validation, similar to the tuning of regularization parameters in regularization-based graph learning algorithms (e.g., Ghoshal and Honorio, 2018; Chen et al., 2019; Park and Kim, 2021). Additionally, data-adaptive procedures with error rate control, such as those proposed by Dai et al. (2023); Guo et al. (2023); Du et al. (2023), can be applied for significance level selection.

We conclude with several open questions. A significant computational cost has been incurred due to the trade-off between computational and sample complexities in the proposed algorithm. Hence, it would be beneficial to develop computationally efficient algorithms while maintaining comparable sample efficiency. Moreover, generalizing these results to non-sub-Gaussian LiNGAMs is of interest. In cases involving independence tests based on distance covariance, the proposed algorithm can recover the true structure with a suffi-



(a) True Graph (b) OptLiNGAM (c) DirectLiNGAM

Figure 3: True and estimated status attainment graphs by OptLiNGAM and DirectLiNGAM algorithms.

ciently large sample size, provided that all error distributions have a finite first moment (Székely et al., 2007). Additionally, any suitable off-the-shelf independence test can be employed while ensuring the consistency of the proposed algorithm. It is conjectured that the proposed algorithm could remain consistent and potentially optimal, even with heavy tailed error distributions.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korea government(MSIT) (NRF-2021R1C1C1004562 and RS-2023-00218231), Institute of Information & communications Technology Planning & Evaluation (IITP) Grant funded by the Korea government(MSIT) (No. 2021-0-01343), and LAMP Program of the National Research Foundation of Korea(NRF) grant funded by the Ministry of Education(No. RS-2023-00301976). Finally, thanks to Professor Shimizu Shohei for his valuable comments.

References

Chen, W., Drton, M., and Wang, Y. S. (2019). On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023). A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association*, 118(543):1551–1565.

Darmois, G. (1953). Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8.

Du, L., Guo, X., Sun, W., and Zou, C. (2023). False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association*, 118(541):607–621.

Duncan, O. D., Featherman, D. L., and Duncan, B. (1972). Socioeconomic background and achievement. Seminar Press.

Eberhardt, F. (2017). Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3:81–91.

Eriksson, J. and Koivunen, V. (2004). Identifiability, separability, and uniqueness of linear ica models. *IEEE signal processing letters*, 11(7):601–604.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gao, M., Tai, W. M., and Aragam, B. (2022). Optimal estimation of gaussian dag models. In *International Conference on Artificial Intelligence and Statistics*, pages 8738–8757. PMLR.

Ghoshal, A. and Honorio, J. (2017). Information-theoretic limits of bayesian network structure learning. In *Artificial Intelligence and Statistics*, pages 767–775. PMLR.

Ghoshal, A. and Honorio, J. (2018). Learning linear structural equation models in polynomial time and sample complexity. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524.

Gong, M., Zhang, K., Schölkopf, B., Glymour, C., and Tao, D. (2017). Causal discovery from temporally aggregated time series. In *Uncertainty in artificial intelligence: proceedings of the... conference. Con-*

ference on Uncertainty in Artificial Intelligence, volume 2017. NIH Public Access.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. *Advances in neural information processing systems*, 20.

Guo, X., Ren, H., Zou, C., and Li, R. (2023). Threshold selection in feature screening for error rate control. *Journal of the American Statistical Association*, 118(543):1773–1785.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

Huo, X. and Székely, G. J. (2016). Fast computing for distance covariance. *Technometrics*, 58(4):435–447.

Hwang, S., Lee, K., Oh, S., and Park, G. (2023). Bayesian approach to linear bayesian networks. *arXiv preprint arXiv:2311.15610*.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *The Journal of Machine Learning Research*, 14(1):111–152.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105.

Maeda, T. N. and Shimizu, S. (2020). Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pages 735–745. PMLR.

Park, G. (2020). Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34.

Park, G. (2023). Computationally efficient learning of gaussian linear structural equation models with equal error variances. *Journal of Computational and Graphical Statistics*, 32(3):1060–1073.

Park, G. and Kim, Y. (2021). Learning high-dimensional gaussian linear structural equation models with heterogeneous error variances. *Computational Statistics & Data Analysis*, 154:107084.

Park, G., Moon, S. J., Park, S., and Jeon, J.-J. (2021). Learning a high-dimensional linear structural equation model via l1-regularized regression. *Journal of Machine Learning Research*, 22(102):1–41.

Pearl, J. et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3.

Peters, J. and Bühlmann, P. (2014). Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ 1-penalized log-determinant divergence.

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

Shimizu, S. and Hyvärinen, A. (2007). Discovery of linear non-gaussian acyclic models in the presence of latent classes. In *International Conference on Neural Information Processing*, pages 752–761. Springer.

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

Skitovitch, V. P. (1953). On a property of the normal distribution. *DAN SSSR*, 89:217–219.

Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances.

Tramontano, D., Monod, A., and Drton, M. (2022). Learning linear non-gaussian polytree models. In *Uncertainty in Artificial Intelligence*, pages 1960–1969. PMLR.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, Y. S. and Drton, M. (2020). High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1):41–59.

Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceed-*

ings of the twenty-fifth conference on uncertainty in artificial intelligence, pages 647–655. AUAI Press.

Zhao, R., He, X., and Wang, J. (2022). Learning linear non-gaussian directed acyclic graph with diverging number of nodes. *Journal of Machine Learning Research*, 23(269):1–34.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/**No**/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/**No**/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/**No**/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Supplementary: Optimal estimation of linear non-Gaussian structure equation models

Sunmin Oh[1] **Seungsu Han**[1] **Gunwoong Park**[1,2,3]
[1]Department of Statistics, Seoul National University, Republic of Korea
[2]Institute for Data Innovation in Science, Seoul National University, Republic of Korea
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Republic of Korea

8 Computational Complexity of Algorithm 1

This section discusses the computational complexity of Algorithm 1 when the (residualized) random variable estimators $\hat{e}_{j,C}$ and $\hat{r}_C(j,k)$ are obtained from the sample covariance matrix, as defined in Equation (4). Although the proposed algorithm allows multiple steps per iteration for ordering estimation, the computational cost is analyzed under the worst-case scenario, where the ordering is inferred sequentially, one node at a time. Furthermore, we assume that all the independence test results are correct, so the size of conditioning set q is limited to d_{in}.

Lemma 18. *Consider a LiNGAM. Suppose that the independence test based on distance covariance is employed and all the independence test results during Algorithm 1 are correct. Then, the worst-case computational complexity of Algorithm 1 is*

$$O(n\log(n)p^{d_{in}+1}d_{in}^4(p-d_{in})).$$

Proof. We outline the computational complexities given the sample covariance matrix, which requires $O(np^2)$ time to compute.

The computational complexity of Step 1) using the distance covariance measure (Huo and Székely, 2016) is of order $O(p^2 n\log(n))$. In terms of Step 2), prior to performing independence test, additional computations are needed to obtain $\hat{e}_{j,C}$ and $\hat{r}_C(j,k)$. In the r-th iteration, calculating $\hat{e}_{j,C}$ and $\hat{r}_C(j,k)$ takes $O((\min\{d_{in}, r-1\}+1)^3)$ and $O((\min\{d_{in}, r-1\}+2)^3)$. Additionally, $p-1$ independence tests are required for each $j \in \{\pi_r, ..., \pi_p\}$ and $C \subset \{\pi_1, ..., \pi_{r-1}\}$ with $|C| \le d_{in}$. Hence, the computational complexity of Step 2) is

$$\begin{aligned}
&\sum_{r=2}^{p} O\left((p-1)\times n\log(n)\times(p-r+1)\sum_{t=1}^{\min(d_{in},r-1)}\binom{r-1}{t}\times(\min\{d_{in},r-1\}+2)^3\right)\\
&= \sum_{r=2}^{d_{in}+1} O\left(p(p-1)\times n\log(n)\times 2^{r-1}\times(r+1)^3\right)\\
&\quad + \sum_{r=d_{in}+2}^{p} O\left(p(p-d_{in}-1)\times n\log(n)\times\sum_{t=1}^{d_{in}}\binom{r-1}{t}\times(d_{in}+2)^3\right)\\
&= O\left(p^2\times n\log(n)\times d_{in}^4 2^{d_{in}}\right) + O\left(p(p-d_{in})\times n\log(n)\times d_{in}^4 p^{d_{in}}\right).
\end{aligned}$$

Consequently, the worst-case computational complexity of Algorithm 1 is

$$\begin{aligned}
&O(np^2) + O(np^2\log(n)) + O\left(p^2\times n\log(n)\times d_{in}^4 2^{d_{in}}\right) + O\left(p(p-d_{in})\times n\log(n)\times d_{in}^4 p^{d_{in}}\right)\\
&= O(n\log(n)p^{d_{in}+1}d_{in}^4(p-d_{in})).
\end{aligned}$$

□

The computational complexity of Algorithm 1 highly depends on the maximum indegree d_{in}, and hence, making it more computationally efficient when d_{in} is small. Additionally, if every non-source node has the same indegree, the ordering estimated by Algorithm 1 matches the topological layers, and the number of iterations required in Step 2) is $T-1$, where the true graph structure has T topological layers. Consequently, under the equal indegree setting, the proposed algorithm is more computationally efficient, and even when the number of topological layers T is small, it achieves further improved computational efficiency.

Unfortunately, if errors occur during independence testing, the computational complexity can increase exponentially with the number of nodes p. Hence, it is practical to limit the size of the conditioning sets. As long as this bound is set to be equal to or larger than d_{in}, the proposed algorithm remains consistent.

9 Proof for Proposition 1

We first introduce Darmois-Skitovitch theorem introduced in (Darmois, 1953; Skitovitch, 1953), which plays a key role to show the following conditional independence properties of LiNGAMs.

Lemma 8. *(Darmois-Skitovitch theorem) Define two random variables u_1 and u_2 as linear combinations of independent random variables $s_i, i = 1, \dots, m$, such that*

$$u_1 = \sum_{i=1}^{m} c_{1,i} s_i \quad \text{and} \quad u_2 = \sum_{i=1}^{m} c_{2,i} s_i.$$

If u_1 and u_2 are independent, all variables s_i with $c_{1,i}c_{2,i} \neq 0$ are Gaussian distributed.

Proposition 1. *Let $P(X)$ be generated from a LiNGAM with DAG G and the topological layers $\cup_{t=0}^{T-1}\mathcal{A}_t$. For any $r \in \{1, 2, ..., T-1\}$, $j \in \mathcal{A}_r$, $k \in V \setminus \cup_{t=0}^{r}\mathcal{A}_t$, and $S_{r-1} = \cup_{t=0}^{r-1}\mathcal{A}_t$,*

(i) there exists $C \subset S_{r-1}$ satisfying $Pa(j) \subset C$, and then $e_{j,C} \perp\!\!\!\perp r_C(\ell, j)$ for all $\ell \in V \setminus (S_{r-1} \cup \{j\})$.

(ii) if $C \subset S_{r-1}$ satisfies $e_{j,C} \perp\!\!\!\perp X_\ell$ for all $\ell \in S_{r-1}$, then $Pa(j) \subset C$.

(iii) for any $C \subset S_{r-1}$, there exists $\ell \in V \setminus (S_{r-1} \cup \{k\})$ such that $e_{k,C} \not\perp\!\!\!\perp r_C(\ell, k)$.

Proof. Let $B \in \mathbb{R}^{p\times p}$ be the edge weight matrix for each element, $B_{j,k} = \beta_{k,j}$. Denote $X_j = \sum_{h\in V} b_{h,j}\epsilon_h$, where $b_{h,j}$ is the (j,h)-th element of $(I-B)^{-1}$. Additionally, for a small C, denote $e_{j,C} = \sum_{h\in V} a_{h,j}\epsilon_h$, where $a_{h,j}$ is the (j,h)-th element of $A = (I-B)^{-1} - \Sigma_{V,C}\Sigma_{C,C}^{-1}[(I-B)^{-1}]_{C,V}$.

Using the path interpretation, the (j,h)-th element of $(I-B)^{-1}$ represents the cumulative influence, $\beta_{h\to j}$, along with all directed paths from h to j, where each path's contribution is the product of the coefficients $(\beta_{h,j})$ along the path. Hence, $b_{h,j} = 0$ if $h \notin \mathrm{An}(j)$ and $b_{h,j} = 1$ if $h = j$. However, this is not necessarily true for $a_{h,j}$. Furthermore, $b_{h,j}$ can be zero even when $h \in \mathrm{Pa}(j)$, if the faithfulness assumption is violated, and similarly, $a_{h,j}$ can be zero even when $h \in \mathrm{Pa}(j)$ and $C \subset \mathrm{Nd}(j)$, violating the parental faithfulness assumption. Importantly, neither type of faithfulness assumption is required in this context.

To prove (i), take $C \subset S_{r-1}$ satisfying $\mathrm{Pa}(j) \subset C$, which is guaranteed by the fact that $\mathrm{Pa}(j) \subset \mathrm{An}(j) \subset S_{r-1}$. Then, from the definition of the topological layers, $\mathrm{Pa}(j) \subset C \subset \mathrm{Nd}(j)$, and hence, $e_{j,C} = \epsilon_j$. Since

$$r_C(\ell, j) = e_{\ell,C} - \frac{\mathrm{Cov}(e_{\ell,C}, \epsilon_j)}{\mathrm{Var}(\epsilon_j)}\epsilon_j = \sum_{h\in V} a_{h,\ell}\epsilon_h - a_{j,\ell}\epsilon_j = \sum_{h\in V\setminus\{j\}} a_{h,\ell}\epsilon_h,$$

the proof follows directly from the independence properties of the error terms.

Now we prove (ii). If $C = \emptyset$, $e_{j,C} = X_j$ and $X_j \not\perp\!\!\!\perp X_\ell$ for $\ell \in \mathrm{Pa}(j) \subset S_{r-1}$, which results in a contradiction. Thus we consider the case of $C \neq \emptyset$. For simplicity in notation, let $\Sigma_{j|C} = \Sigma_{j,j} - \Sigma_{j,C}\Sigma_{C,C}^{-1}\Sigma_{C,j}$, where $j \in V$

and $C \subset V \setminus \{j\}$. Note that, if $[\Sigma_{j,C\cup\{c\}}(\Sigma_{C\cup\{c\},C\cup\{c\}})^{-1}]_c = 0$ for $j, c \in V$ and $C \subset V \setminus \{j,c\}$,

$$\begin{aligned}
&\Sigma_{j,C\cup\{c\}}(\Sigma_{C\cup\{c\},C\cup\{c\}})^{-1} \\
&= \begin{bmatrix}\Sigma_{j,c} & \Sigma_{j,C}\end{bmatrix}\begin{bmatrix}\Sigma_{c,c} & \Sigma_{c,C} \\ \Sigma_{C,c} & \Sigma_{C,C}\end{bmatrix}^{-1} \\
&= \begin{bmatrix}\Sigma_{j,c} & \Sigma_{j,C}\end{bmatrix}\begin{bmatrix}\Sigma_{c|C}^{-1} & -\Sigma_{c|C}^{-1}\Sigma_{c,C}\Sigma_{C,C}^{-1} \\ -\Sigma_{C,C}^{-1}\Sigma_{C,c}\Sigma_{c|C}^{-1} & \Sigma_{C,C}^{-1} + \Sigma_{C,C}^{-1}\Sigma_{C,c}\Sigma_{c|C}^{-1}\Sigma_{c,C}\Sigma_{C,C}^{-1}\end{bmatrix} \\
&= \begin{bmatrix}\Sigma_{j,c}\Sigma_{c|C}^{-1} - \Sigma_{j,C}\Sigma_{C,C}^{-1}\Sigma_{C,c}\Sigma_{c|C}^{-1} & -\Sigma_{j,c}\Sigma_{c|C}^{-1}\Sigma_{c,C}\Sigma_{C,C}^{-1} + \Sigma_{j,C}\Sigma_{C,C}^{-1}(\Sigma_{C,C} + \Sigma_{C,c}\Sigma_{c|C}^{-1}\Sigma_{c,C})\Sigma_{C,C}^{-1}\end{bmatrix} \\
&= \begin{bmatrix}0 & \Sigma_{j,C}\Sigma_{C,C}^{-1}\end{bmatrix},
\end{aligned}$$

and hence, $e_{j,C\cup\{c\}} = e_{j,C}$. Thus we restrict our attention to the case where all elements of $\Sigma_{j,C}\Sigma_{C,C}^{-1}$ are nonzero.

Suppose $S_{r-1} \setminus \mathrm{An}(j) \neq \emptyset$ and take $\ell \in S_{r-1} \setminus \mathrm{An}(j)$ satisfying $\mathrm{De}(\ell) \cap S_{r-1} = \emptyset$ by the acyclicity assumption. In this case, we have $b_{\ell,u} = 0$ for all $u \in S_{r-1} \setminus \{\ell\}$ and $b_{\ell,\ell} = 1$. Consequently, if $\ell \in C$, Lemma 8 asserts that $e_{j,C} \not\perp\!\!\!\perp X_\ell$. Equivalently, $e_{j,C} \perp\!\!\!\perp X_\ell$ implies $\ell \notin C$. Similarly, take $\ell' \in S_{r-1} \setminus (\mathrm{An}(j) \cup \{\ell\})$ satisfying $\mathrm{De}(\ell') \cap S_{r-1} \setminus \{\ell\} = \emptyset$. Then $b_{\ell',u} = 0$ for $u \in S_{r-1} \setminus \{\ell, \ell'\}$ and $b_{\ell',\ell'} = 1$. By applying Lemma 8, we deduce $\ell' \notin C$, and hence, by iterating this procedure, we conclude that $C \subset \mathrm{An}(j)$.

The remainder of the proof is established by employing a contrapositive argument. Note that $e_{j,C}$ is represented as

$$\begin{aligned}
e_{j,C} &= \sum_{h\in\mathrm{An}(j)} b_{h,j}\epsilon_h + \epsilon_j - \Sigma_{j,C}\Sigma_{C,C}^{-1}[(I-B)^{-1}]_{C,V}\epsilon \\
&= \sum_{h\in\mathrm{An}(j)} b_{h,j}\epsilon_h + \epsilon_j - \Sigma_{j,C}\Sigma_{C,C}^{-1}\left[\sum_{h\in\mathrm{An}(m)} b_{h,m}\epsilon_h + \epsilon_m\right]_{m\in C}.
\end{aligned}$$

Suppose $\mathrm{Pa}(j) \not\subset C \subset \mathrm{An}(j)$. Then, there exists an $\ell \in \mathrm{Pa}(j) \setminus C$, and by Lemma 8, $e_{j,C} \perp\!\!\!\perp X_\ell$ implies that $b_{\ell,j} = \Sigma_{j,C}\Sigma_{C,C}^{-1}[b_{\ell,m}]_{m\in C}$.

If $b_{\ell,j} = \Sigma_{j,C}\Sigma_{C,C}^{-1}[b_{\ell,m}]_{m\in C} \neq 0$, then $C \cap \mathrm{De}(\ell) \neq \emptyset$. Conversely, when $b_{\ell,j} = 0$, this indicates a path cancellation, implying that $\mathrm{An}(j) \cap \mathrm{De}(\ell) \neq \emptyset$. Given that $C \subset \mathrm{An}(j)$, in both cases of $b_{\ell,j} \neq 0$ and $b_{\ell,j} = 0$, we can take a node $m_1 \in \mathrm{An}(j) \cap \mathrm{De}(\ell) \subset S_{r-1}$. By a similar argument, $e_{j,C} \perp\!\!\!\perp X_{m_1}$ implies that $b_{m_1,j} = \Sigma_{j,C}\Sigma_{C,C}^{-1}[b_{m_1,m}]_{m\in C}$, allowing us to choose a node $m_2 \in \mathrm{An}(j) \cap \mathrm{De}(m_1)$. Repeating this process iteratively, we can construct a set $\{m_i : m_i \in \mathrm{De}(\ell), m_{i+1} \in \mathrm{De}(m_i), i = 1, 2, ...\} \subset \mathrm{An}(j)$. By the acyclicity assumption, each m_i is distinct, which implies the construction of an infinite set $\{m_i : m_1 \in \mathrm{De}(\ell), m_{i+1} \in \mathrm{De}(m_i), i = 1, 2, ...\}$ contained within $\mathrm{An}(j)$. This leads to a contradiction.

Finally we prove *(iii)*. Let $k \in \mathcal{A}_s$ for $r < s \leq T-1$ and $\ell \in (\mathrm{Pa}(k) \cap \mathcal{A}_{s-1}) \subset V \setminus (S_{r-1} \cup \{k\})$. Then, for any $C \subset S_{r-1}$, we have $(\{\ell\} \cup \mathrm{De}(\ell)) \cap C = \emptyset$, $a_{\ell,k} = b_{\ell,k}$ and $a_{k,k} = a_{\ell,\ell} = 1$. Furthermore, since $k \in \mathcal{A}_s$ and $\ell \in \mathcal{A}_{s-1}$, there is only one path from ℓ to k, and hence, $b_{\ell,k} \neq 0$. Note that $r_C(\ell,k)$ is represented as

$$\begin{aligned}
r_C(\ell,k) &= e_{\ell,C} - \frac{\mathrm{Cov}(e_{\ell,C}, e_{k,C})}{\mathrm{Var}(e_{k,C})}e_{k,C} \\
&= \left(1 - \frac{\mathrm{Cov}(e_{\ell,C}, e_{k,C})}{\mathrm{Var}(e_{k,C})}a_{\ell,k}\right)\epsilon_\ell - \frac{\mathrm{Cov}(e_{\ell,C}, e_{k,C})}{\mathrm{Var}(e_{k,C})}\epsilon_k + \sum_{h\in V\setminus\{\ell,k\}}\left(a_{h,\ell} - \frac{\mathrm{Cov}(e_{\ell,C}, e_{k,C})}{\mathrm{Var}(e_{k,C})}a_{h,k}\right)\epsilon_h.
\end{aligned}$$

If $\frac{\mathrm{Cov}(e_{\ell,C},e_{k,C})}{\mathrm{Var}(e_{k,C})} = 0$, $a_{\ell,k}\left(1 - \frac{\mathrm{Cov}(e_{\ell,C},e_{k,C})}{\mathrm{Var}(e_{k,C})}a_{\ell,k}\right) = a_{\ell,k} \neq 0$, and by Lemma 8, we have $e_{k,C} \not\perp\!\!\!\perp r_C(\ell,k)$. On the other hand, if $\frac{\mathrm{Cov}(e_{\ell,C},e_{k,C})}{\mathrm{Var}(e_{k,C})} \neq 0$, then $a_{k,k}\frac{\mathrm{Cov}(e_{\ell,C},e_{k,C})}{\mathrm{Var}(e_{k,C})} \neq 0$, and by Lemma 8, we again conclude that $e_{k,C} \not\perp\!\!\!\perp r_C(\ell,k)$.

□

10 Proof for Theorem 2

Theorem 2. *Let $P(X)$ be generated from a LiNGAM with DAG G and the topological layers $\cup_{t=0}^{T-1}\mathcal{A}_t = V$. Consider any $r \in \{1, 2, ..., T-1\}$ and $S_{r-1} = \cup_{t=0}^{r-1}\mathcal{A}_t$. Then*

$$\begin{aligned}\mathcal{A}_0 &= \left\{j \in V : X_j \perp\!\!\!\perp e_{k,\{j\}} \text{ for all } k \in V \setminus \{j\}\right\}, \text{ and}\\ \mathcal{A}_r &= \{j \in V \setminus S_{r-1} : \exists C_j \subset S_{r-1} \text{ s.t. } X_k \perp\!\!\!\perp e_{j,C_j} \text{ for all } k \in S_{r-1} \text{ and}\\ &\qquad e_{j,C_j} \perp\!\!\!\perp r_{C_j}(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\})\}.\end{aligned} \tag{2}$$

Furthermore, for each $j \in \mathcal{A}_r$ and the corresponding set

$$C_j = \{C \subset S_{r-1} : X_k \perp\!\!\!\perp e_{j,C} \text{ for all } k \in S_{r-1} \text{ and } e_{j,C} \perp\!\!\!\perp r_C(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\})\},$$

$Pa(j) \subset C \subset Nd(j)$ for any $C \in C_j$.

Proof. The process of recovering the first topological layer $\mathcal{A}_0$ follows the method outlined by Shimizu et al. (2011). Hence, we omit the proof. The remaining steps are derived directly from Proposition 1.

Specifically, given S_{r-1} for any $r \in \{1, 2, ..., T-1\}$, suppose $j \in \mathcal{A}_r$. By definition of topological layers, we can take a set $C_j \subset S_{r-1}$ satisfying $\mathrm{Pa}(j) \subset C_j \subset \mathrm{Nd}(j)$. Then, $e_{j,C_j} = \epsilon_j$ and $X_k \perp\!\!\!\perp e_{j,C_j}$ for all $k \in S_{r-1}$. Additionally, by Proposition 1 (i), $e_{j,C_j} \perp\!\!\!\perp r_{C_j}(\ell, j)$ for all $\ell \in V \setminus (S_{r-1} \cup \{j\})$. Hence, we have

$$\begin{aligned}\mathcal{A}_r \subset \{j \in V \setminus S_{r-1} : \exists C_j \subset S_{r-1} \text{ s.t. } X_k \perp\!\!\!\perp e_{j,C_j} \text{ for all } k \in S_{r-1} \text{ and}\\ e_{j,C_j} \perp\!\!\!\perp r_{C_j}(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\})\}.\end{aligned}$$

Now, let $k \in V \setminus \cup_{t=0}^{r}\mathcal{A}_t$. Then, by Proposition 1 (iii), for any $C \subset S_{r-1}$ there exists $\ell \in V \setminus (S_{r-1} \cup \{k\})$ such that $e_{k,C} \not\perp\!\!\!\perp r_C(\ell, k)$, and hence, we have

$$\begin{aligned}k \notin \{j \in V \setminus S_{r-1} : \exists C_j \subset S_{r-1} \text{ s.t. } X_k \perp\!\!\!\perp e_{j,C_j} \text{ for all } k \in S_{r-1} \text{ and}\\ e_{j,C_j} \perp\!\!\!\perp r_{C_j}(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\})\}.\end{aligned}$$

Therefore, Equation (2) holds.

Finally, for each $j \in \mathcal{A}_r$, denote the corresponding set as

$$C_j = \{C \subset S_{r-1} : X_k \perp\!\!\!\perp e_{j,C} \text{ for all } k \in S_{r-1} \text{ and } e_{j,C} \perp\!\!\!\perp r_C(\ell, j) \text{ for all } \ell \in V \setminus (S_{r-1} \cup \{j\})\}.$$

Then, for any $j \in V$ and $C \in C_j$, we have $\mathrm{Pa}(j) \subset C \subset \mathrm{Nd}(j)$ by Proposition 1 (ii). □

11 Proof for Theorem 5

Theorem 5. *Consider a sub-Gaussian LiNGAM with $d_{in} \le \frac{p}{2}$. Suppose that Assumptions 3 and 4 are satisfied. Let $\widehat{G} = (V, \widehat{E})$ be the estimated graph from Algorithm 1 with the significance level of the independence test $\alpha_n = 2(1 - \Phi(\sqrt{n\epsilon_0}))$ for any $\epsilon_0 \in (0, \tau_1/2)$. Then, there exist positive constants D_1 and D_2 such that*

$$\Pr(\widehat{G} = G) \ge 1 - D_1\left(\frac{p}{d_{in}}\right)^{d_{in}} exp\left(-\frac{D_2\epsilon_0^4}{\lambda^{36}}n\right).$$

Proof. As discussed in Section 4, the topological ordering and parent sets of each node are correctly recovered as long as no type I and type II errors occur during the process. Additionally, if all the independence tests produce correct results, the conditioning sets evaluated throughout the procedure will be of size less than or equal to d_{in}.

Denote $\mathcal{E}_{jk,C} = \mathcal{E}^I_{jk,C} \cup \mathcal{E}^{II}_{jk,C}$, where

$$\begin{aligned}\text{Type } I \text{ error} : \mathcal{E}^I_{jk,C} &= \left\{n\widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k) > (\Phi^{-1}(1-\alpha/2))^2 \text{ and } T(e_{j,C}, X_k) = 0\right\},\\ \text{Type } II \text{ error} : \mathcal{E}^{II}_{jk,C} &= \left\{n\widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k) \le (\Phi^{-1}(1-\alpha/2))^2 \text{ and } T(e_{j,C}, X_k) > 0\right\}.\end{aligned}$$

Similarly, denote $\widetilde{\mathcal{E}}_{j\ell,C} = \widetilde{\mathcal{E}}^{I}_{j\ell,C} \cup \widetilde{\mathcal{E}}^{II}_{j\ell,C}$, where

$$\text{Type } I \text{ error} : \widetilde{\mathcal{E}}^{I}_{j\ell,C} = \left\{ n\widehat{T}(\hat{e}_{j,C}, \hat{r}_C(\ell,j)) > (\Phi^{-1}(1-\alpha/2))^2 \text{ and } T(e_{j,C}, r_C(\ell,j)) = 0 \right\},$$
$$\text{Type } II \text{ error} : \widetilde{\mathcal{E}}^{II}_{j\ell,C} = \left\{ n\widehat{T}(\hat{e}_{j,C}, \hat{r}_C(\ell,j)) \leq (\Phi^{-1}(1-\alpha/2))^2 \text{ and } T(e_{j,C}, r_C(\ell,j)) > 0 \right\}.$$

Then, we obtain the following probability bounds.

$$\Pr(\mathcal{E}^{I}_{jk,C}) \leq \Pr\left(|T(e_{j,C}, X_k) - \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k)| > \frac{1}{n}(\Phi^{-1}(1-\alpha/2))^2 \right),$$
$$\Pr(\mathcal{E}^{II}_{jk,C}) \leq \Pr\left(|T(e_{j,C}, X_k) - \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k)| \geq T(e_{j,C}, X_k) - \frac{1}{n}(\Phi^{-1}(1-\alpha/2))^2 \right).$$

Following a similar treatment to $\widetilde{\mathcal{E}}_{j\ell,C}$ and setting $\alpha = 2(1-\Phi(\sqrt{n\epsilon_0}))$ for any $\epsilon_0 \in (0, \tau_1/2)$,

$$\Pr(\hat{G} \neq G) \leq \Pr\left(\left(\bigcup_{j\in V} \bigcup_{k\in V\setminus\{j\}} \bigcup_{\substack{C\subset V\setminus\{j\} \\ |C|\leq d_{in}}} \mathcal{E}_{jk,C} \right) \bigcup \left(\bigcup_{j\in V} \bigcup_{\ell\in V\setminus\{j\}} \bigcup_{\substack{C\subset V\setminus\{j,\ell\} \\ |C|\leq d_{in}}} \widetilde{\mathcal{E}}_{j\ell,C} \right) \right)$$
$$\leq \underbrace{\Pr\left(\bigcup_{\substack{j\neq k\in V \\ C\subset V\setminus\{j\} \\ |C|\leq d_{in}}} \left\{ |T(e_{j,C}, X_k) - \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k)| > \epsilon_0 \right\} \right)}_{\mathcal{P}_1}$$
$$+ \underbrace{\Pr\left(\bigcup_{\substack{j\neq \ell\in V \\ C\subset V\setminus\{j,\ell\} \\ |C|\leq d_{in}}} \left\{ |T(e_{j,C}, r_C(\ell,j)) - \widehat{T}(\hat{e}_{j,C}, \hat{r}_C(\ell,j))| > \epsilon_0 \right\} \right)}_{\mathcal{P}_2}.$$

Applying the union bound, we have the upper bounds of P_1 and P_2 as follows.

$$\mathcal{P}_1 \leq \sum_{j\in V} \sum_{k\in V\setminus\{j\}} \sum_{\substack{C\subset V\setminus\{j\} \\ |C|\leq d_{in}}} \Pr\left(|T(e_{j,C}, X_k) - \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k)| \geq \epsilon_0 \right)$$
$$\leq \max_{\substack{j,k\in V \\ C\subset V\setminus\{j\} \\ |C|\leq d_{in}}} \Pr\left(|T(e_{j,C}, X_k) - \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k)| \geq \epsilon_0 \right) \times p(p-1) \times \sum_{q=0}^{d_{in}} \binom{p-1}{q},$$
$$\mathcal{P}_2 \leq \sum_{j\in V} \sum_{\ell\in V\setminus\{j\}} \sum_{\substack{C\subset V\setminus\{j,\ell\} \\ |C|\leq d_{in}}} \Pr\left(|T(e_{j,C}, r_C(\ell,j)) - \widehat{T}(\hat{e}_{j,C}, \hat{r}_C(\ell,j))| \geq \epsilon_0 \right)$$
$$\leq \max_{\substack{j,\ell\in V \\ C\subset V\setminus\{j,\ell\} \\ |C|\leq d_{in}}} \Pr\left(|T(e_{j,C}, r_C(\ell,j)) - \widehat{T}(\hat{e}_{j,C}, \hat{r}_C(\ell,j))| \geq \epsilon_0 \right) \times p(p-1) \times \sum_{q=0}^{d_{in}} \binom{p-2}{q}.$$

Since $\binom{n}{k} < \left(\frac{ne}{k}\right)^k$ for all values of n and k such that $1 \leq k \leq n$ and $d_{in} < \frac{p}{2}$, we have

$$\max\left\{ \sum_{q=0}^{d_{in}} \binom{p-1}{q}, \sum_{q=0}^{d_{in}} \binom{p-2}{q} \right\} \leq (d_{in}+1) \times \exp(d_{in}) \times \exp\left((d_{in}) \log \frac{p}{d_{in}} \right). \tag{5}$$

Now, we introduce two lemmas representing error bounds of test statistic estimator.

Lemma 9. *Suppose that Assumptions 3 and 4 are satisfied. Consider the test statistic T for independence, as defined in Equation (6). For any $j \in V$, $C \subset V \setminus \{j\}$, and $\epsilon > 0$, there exists a positive constant K_1 such that*

$$\Pr\left(|T(e_{j,C}, X_k) - \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k)| \geq \epsilon\right) = \mathcal{O}\left(exp\left(-\frac{K_1\epsilon^4}{\lambda^{16}}n + 4(|C|+2)\right)\right).$$

Lemma 10. *Suppose that Assumptions 3 and 4 are satisfied. Consider the test statistic T for independence, as in Equation (9). For any $j, \ell \in V$, set $C \subset V \setminus \{j, \ell\}$, and $\epsilon > 0$, there exists a positive constant K_2 such that*

$$\Pr\left(|T(e_{j,C}, r_C(\ell, j)) - \widehat{T}(\hat{e}_{j,C}, \hat{r}_C(\ell, j))| \geq \epsilon\right) = \mathcal{O}\left(exp\left(-\frac{K_2\epsilon^4}{\lambda^{36}}n + 4(|C|+2)\right)\right).$$

Consequently, by combining Lemma 9, Lemma 10, and Equation (5), the proof of Theorem 5 is complete.

$$\begin{aligned}
\Pr(\hat{G} \neq G) &= \mathcal{O}\left(\exp\left(-\frac{K_1\epsilon^4}{\lambda^{16}}n + 4(d_{in}+2)\right) \times \exp\left(\log(d_{in}+1) + (d_{in})\left(1 + \log\frac{p}{d_{in}}\right)\right)\right) \\
&\quad + \mathcal{O}\left(\exp\left(-\frac{K_2\epsilon^4}{\lambda^{36}}n + 4(d_{in}+2)\right) \times \exp\left(\log(d_{in}+1) + (d_{in})\left(1 + \log\frac{p}{d_{in}}\right)\right)\right) \\
&= \mathcal{O}\left(\exp\left(d_{in}\log\left(\frac{p}{d_{in}}\right) - \frac{K_1 \wedge K_2\epsilon^4}{\lambda^{36}}n\right)\right).
\end{aligned}$$

□

11.1 Notations

Before providing necessary lemmas and proofs, we first provide some notations. For any $j \in V$ and a small $C \subset V \setminus \{j\}$, $\widetilde{e}_{j,C} = \mathbf{x}_j - \mathbf{x}_C(\Sigma_{C,C})^{-1}\Sigma_{C,j}$. For simplicity in notation, let $\Sigma_{j,k|C} = \Sigma_{j,k} - \Sigma_{j,C}(\Sigma_{C,C})^{-1}\Sigma_{C,k}$ for any $j, k \in V$ and $C \subset V \setminus \{j,k\}$. Additionally, we denote $\Sigma_{j|C} = \Sigma_{j,j|C}$. The same notation is applied to $\widehat{\Sigma}$ as well. Then, for any $j, k \in V$ and a small $C \subset V \setminus \{j,k\}$, $\hat{r}_C(j,k) = \hat{e}_{j,C} - \frac{\widehat{\Sigma}_{j,k|C}}{\widehat{\Sigma}_{k|C}}\hat{e}_{k,C}$, and $\widetilde{r}_C(j,k) = \widetilde{e}_{j,C} - \frac{\Sigma_{j,k|C}}{\Sigma_{k|C}}\widetilde{e}_{k,C}$. Lastly, $\|\cdot\|_2$ and $\|\cdot\|$ denote the ℓ_2-norm of a matrix and a vector, respectively.

Furthermore, the considered independence test statistic and its sample and estimated versions are defined as follows. For any $j \in V$ and $C \subset V \setminus \{j\}$,

$$T(e_{j,C}, X_k) = \frac{\mathrm{dcov}^2(e_{j,C}, X_k)}{I_{jk,2}},\ \widehat{T}(\widetilde{e}_{j,C}, \mathbf{x}_k) = \frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \mathbf{x}_k)}{\widetilde{I}_{jk,2}},\ \text{and}\ \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k) = \frac{\widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \mathbf{x}_k)}{\widehat{I}_{jk,2}}. \tag{6}$$

Here, for any set $S \subset V \setminus \{j\}$, $\mathrm{dcov}^2(e_{j,C}, X_S) = I_{jS,1} + I_{jS,2} - 2I_{jS,3}$ with

$$\begin{aligned}
I_{jS,1} &= \mathbb{E}[|e_{j,C} - e'_{j,C}|\|X_S - X'_S\|], \\
I_{jS,2} &= \mathbb{E}[|e_{j,C} - e'_{j,C}|]\mathbb{E}[\|X_S - X'_S\|], \\
I_{jS,3} &= \mathbb{E}\left[\mathbb{E}[|e_{j,C} - e'_{j,C}| \mid e_{j,C}]\mathbb{E}[\|X_S - X'_S\| \mid X_S]\right],
\end{aligned}$$

where the notation $'$ denotes independent copy of the corresponding random vector. Additionally,

$$\begin{aligned}
&\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \mathbf{x}_S) = \widetilde{I}_{jS,1} + \widetilde{I}_{jS,2} - 2\widetilde{I}_{jS,3},\ \text{where} \\
&\widetilde{I}_{jS,1} = \frac{1}{n^2}\sum_{i,h=1}^{n} |\widetilde{e}^{(i)}_{j,C} - \widetilde{e}^{(h)}_{j,C}|\|\mathbf{x}^{(i)}_S - \mathbf{x}^{(h)}_S\|, \\
&\widetilde{I}_{jS,2} = \left(\frac{1}{n^2}\sum_{i,h=1}^{n} |\widetilde{e}^{(i)}_{j,C} - \widetilde{e}^{(h)}_{j,C}|\right)\left(\frac{1}{n^2}\sum_{i,h=1}^{n} \|\mathbf{x}^{(i)}_S - \mathbf{x}^{(h)}_S\|\right), \\
&\widetilde{I}_{jS,3} = \frac{1}{n^3}\sum_{i,h,m=1}^{n} |\widetilde{e}^{(i)}_{j,C} - \widetilde{e}^{(m)}_{j,C}|\|\mathbf{x}^{(h)}_S - \mathbf{x}^{(m)}_S\|,
\end{aligned} \tag{7}$$

and

$$\widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \mathbf{x}_S) = \widehat{I}_{jS,1} + \widehat{I}_{jS,2} - 2\widehat{I}_{jS,3}, \text{ where} \tag{8}$$
$$\widehat{I}_{jS,1} = \frac{1}{n^2}\sum_{i,h=1}^{n} |\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(h)}| \|\mathbf{x}_S^{(i)} - \mathbf{x}_S^{(h)}\|,$$
$$\widehat{I}_{jS,2} = \left(\frac{1}{n^2}\sum_{i,h=1}^{n} |\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(h)}|\right)\left(\frac{1}{n^2}\sum_{i,h=1}^{n} \|\mathbf{x}_S^{(i)} - \mathbf{x}_S^{(h)}\|\right),$$
$$\widehat{I}_{jS,3} = \frac{1}{n^3}\sum_{i,h,m=1}^{n} |\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(m)}| \|\mathbf{x}_S^{(h)} - \mathbf{x}_S^{(m)}\|.$$

Similarly, for any $j, \ell \in V$, and $C \subset V \setminus \{j, \ell\}$, we denote

$$T(e_{j,C}, r_C(\ell, j)) = \frac{\mathrm{dcov}^2(e_{j,C}, r_C(\ell, j))}{J_{j\ell,2}}, \tag{9}$$
$$\widehat{T}(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell, j)) = \frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell, j))}{\widetilde{J}_{j\ell,2}},$$
$$\widehat{T}(\hat{e}_{j,C}, \hat{r}_C(\ell, j)) = \frac{\widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \hat{r}_C(\ell, j))}{\widehat{J}_{j\ell,2}}.$$

Here, by simply replacing X_k with $r_C(\ell, j)$, $\mathrm{dcov}^2(e_{j,C}, r_C(\ell, j)) = J_{j\ell,1} + J_{j\ell,2} - 2J_{j\ell,3}$ with

$$J_{j\ell,1} = \mathbb{E}[|e_{j,C} - e'_{j,C}||r_C(\ell, j) - r_C(\ell, j)'|],$$
$$J_{j\ell,2} = \mathbb{E}[|e_{j,C} - e'_{j,C}|]\mathbb{E}[|r_C(\ell, j) - r_C(\ell, j)'|],$$
$$J_{j\ell,3} = \mathbb{E}\left[\mathbb{E}[|e_{j,C} - e'_{j,C}| \mid e_{j,C}]\mathbb{E}[|r_C(\ell, j) - r_C(\ell, j)'| \mid r_C(\ell, j)]\right].$$

Additionally,

$$\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell, j)) = \widetilde{J}_{j\ell,1} + \widetilde{J}_{j\ell,2} - 2\widetilde{J}_{j\ell,3}, \text{ where} \tag{10}$$
$$\widetilde{J}_{j\ell,1} = \frac{1}{n^2}\sum_{i,h=1}^{n} |\widetilde{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(h)}||\widetilde{r}_C(\ell, j)^{(i)} - \widetilde{r}_C(\ell, j)^{(h)}|,$$
$$\widetilde{J}_{j\ell,2} = \left(\frac{1}{n^2}\sum_{i,h=1}^{n} |\widetilde{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(h)}|\right)\left(\frac{1}{n^2}\sum_{i,h=1}^{n} |\widetilde{r}_C(\ell, j)^{(i)} - \widetilde{r}_C(\ell, j)^{(h)}|\right),$$
$$\widetilde{J}_{j\ell,3} = \frac{1}{n^3}\sum_{i,h,m=1}^{n} |\widetilde{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(m)}||\widetilde{r}_C(\ell, j)^{(h)} - \widetilde{r}_C(\ell, j)^{(m)}|,$$

and

$$\widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \widehat{r}_C(\ell, j)) = \widehat{J}_{j\ell,1} + \widehat{J}_{j\ell,2} - 2\widehat{J}_{j\ell,3}, \text{ where} \tag{11}$$
$$\widehat{J}_{j\ell,1} = \frac{1}{n^2}\sum_{i,h=1}^{n} |\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(h)}||\hat{r}_C(\ell, j)^{(i)} - \hat{r}_C(\ell, j)^{(h)}|,$$
$$\widehat{J}_{j\ell,2} = \left(\frac{1}{n^2}\sum_{i,h=1}^{n} |\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(h)}|\right)\left(\frac{1}{n^2}\sum_{i,h=1}^{n} |\hat{r}_C(\ell, j)^{(i)} - \hat{r}_C(\ell, j)^{(h)}|\right),$$
$$\widehat{J}_{j\ell,3} = \frac{1}{n^3}\sum_{i,h,m=1}^{n} |\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(m)}||\hat{r}_C(\ell, j)^{(h)} - \hat{r}_C(\ell, j)^{(m)}|.$$

11.2 Proof for Lemma 9

Proof. Applying the union bound, the error probability can be decomposed as follows.

$$\begin{aligned}
&\Pr\left(|T(e_{j,C}, X_k) - \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k)| \geq \epsilon\right) \\
&\leq \underbrace{\Pr\left(|T(e_{j,C}, X_k) - \widehat{T}(\widetilde{e}_{j,C}, \mathbf{x}_k)| \geq \frac{\epsilon}{2}\right)}_{P_1} + \underbrace{\Pr\left(|\widehat{T}(\widetilde{e}_{j,C}, \mathbf{x}_k) - \widehat{T}(\hat{e}_{j,C}, \mathbf{x}_k)| \geq \frac{\epsilon}{2}\right)}_{P_2} \\
&:= P_1 + P_2.
\end{aligned}$$

Since for any $j \in V$ and $C \subset V \setminus \{j\}$, $e_{j,C}$ and X_k are sub-Gaussian distributed random variables, the condition (C1) in Li et al. (2012) is satisfied for any $s \in \mathbb{R}$. Furthermore, Assumption 4 implies that the condition (C2) in Li et al. (2012) is satisfied with $\kappa = 0$. By following the proof of Theorem 1 in Li et al. (2012), for any $s \in \mathbb{R}$, it holds that

$$P_1 = \mathcal{O}\left(\exp\left(-\left(\frac{\epsilon^2}{M^2} + Ms\right)n\right)\right),$$

where M is a truncation constant. Hence, none of ϵ, M and s depend on n, p or d_{in}.

To establish an upper bound for P_2, it suffices to derive bounds for $\widetilde{I}_{jk,1}, \widetilde{I}_{jk,2}, \widetilde{I}_{jk,3}$ and $\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \mathbf{x}_k)$. Simple algebra yields that

$$\widetilde{I}_{jk,1} \leq \frac{1}{n^2}\sum_{i,h=1}^{n}\left(|\widetilde{e}_{j,C}^{(i)}| + |\widetilde{e}_{j,C}^{(h)}|\right)\left(|\mathbf{x}_k^{(i)}| + |\mathbf{x}_k^{(h)}|\right) = \frac{2}{n}\sum_{i=1}^{n}|\widetilde{e}_{j,C}^{(i)}||\mathbf{x}_k^{(i)}| + 2\left(\frac{1}{n}\sum_{i=1}^{n}|\widetilde{e}_{j,C}^{(i)}|\right)\left(\frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_k^{(i)}|\right). \tag{12}$$

Let $D = \{j,k\} \cup C$. Then, if $\|\widehat{\Sigma}_D - \Sigma_D\|_2 < 1$, Assumption 3 gives that

$$\frac{1}{n}\|\mathbf{x}_k\|^2 = \widehat{\Sigma}_{k,k} \leq \lambda + 1,$$

and

$$\begin{aligned}
\frac{1}{n}\|\widetilde{e}_{j,C}\|^2 &= \widehat{\Sigma}_{j,j} - \Sigma_{j,C}(\Sigma_{C,C})^{-1}\widehat{\Sigma}_{C,C}(\Sigma_{C,C})^{-1}\Sigma_{C,j} \\
&= (\Sigma_{j,j} + 1) - \Sigma_{j,C}(\Sigma_{C,C})^{-1}\Sigma_{C,j} - \Sigma_{j,C}(\Sigma_{C,C})^{-1}(\widehat{\Sigma}_{C,C} - \Sigma_{C,C})(\Sigma_{C,C})^{-1}\Sigma_{C,j} \\
&\leq \Sigma_{j|C} + 1 + \|\Sigma_{j,C}(\Sigma_{C,C})^{-1}\|^2 = 1/[(\Sigma_{\{j\}\cup C,\{j\}\cup C})^{-1}]_{j,j} + \lambda^2 + 1 \leq \lambda^2 + \lambda + 1.
\end{aligned} \tag{13}$$

Hence, applying Jensen's inequality and Hölder's inequality, we have

$$\widetilde{I}_{jk,1} \leq 4\left(\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C}\|\right)\left(\frac{1}{\sqrt{n}}\|\mathbf{x}_k\|\right) \leq 4(\lambda+1)^{3/2}.$$

Additionally, following a similar method, we obtain

$$\widetilde{I}_{jk,2} \leq \left(\frac{1}{n^2}\sum_{i,h=1}^{n}\left(|\widetilde{e}_{j,C}^{(i)}| + |\widetilde{e}_{j,C}^{(h)}|\right)\right)\left(\frac{1}{n^2}\sum_{i,h=1}^{n}\left(|\mathbf{x}_k^{(i)}| + |\mathbf{x}_k^{(h)}|\right)\right) = 4\left(\frac{1}{n}\sum_{i=1}^{n}|\widetilde{e}_{j,C}^{(i)}|\right)\left(\frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_k^{(i)}|\right), \tag{14}$$

$$\widetilde{I}_{jk,3} \leq \frac{1}{n^3}\sum_{i,h,m=1}^{n}\left(|\widetilde{e}_{j,C}^{(i)}| + |\widetilde{e}_{j,C}^{(m)}|\right)\left(|\mathbf{x}_k^{(h)}| + |\mathbf{x}_k^{(m)}|\right) = \frac{1}{n}\sum_{i=1}^{n}|\widetilde{e}_{j,C}^{(i)}||\mathbf{x}_k^{(i)}| + 3\left(\frac{1}{n}\sum_{i=1}^{n}|\widetilde{e}_{j,C}^{(i)}|\right)\left(\frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_k^{(i)}|\right).$$

Consequently, we can conclude that

$$\max\left\{|\widetilde{I}_{jk,1}|, |\widetilde{I}_{jk,2}|, |\widetilde{I}_{jk,3}|\right\} \leq 4(\lambda+1)^{3/2}. \tag{15}$$

Finally, applying Lemma 17, there exist positive constants Q_1 and Q_2 such that

$$\Pr\left(\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C},\mathbf{x}_k)\le 16(\lambda+1)^{3/2}\right)\ge 1-Q_1\exp(-Q_2 n+4(|C|+2)). \tag{16}$$

For any $c<\tau_2=\min_{j\neq k}\widetilde{I}_{jk,2}$, the term P_2 can be bounded as follows.

$$\begin{aligned}
P_2 &\le \Pr\left(|\widehat{T}(\widetilde{e}_{j,C},\mathbf{x}_k)-\widehat{T}(\widehat{e}_{j,C},\mathbf{x}_k)|\ge\frac{\epsilon}{2}\text{ and }|\widehat{I}_{jk,2}|\le c\right)+\Pr\left(|\widehat{T}(\widetilde{e}_{j,C},\mathbf{x}_k)-\widehat{T}(\widehat{e}_{j,C},\mathbf{x}_k)|\ge\frac{\epsilon}{2}\text{ and }|\widehat{I}_{jk,2}|>c\right)\\
&\le \underbrace{\Pr\left(|\widehat{I}_{jk,2}|\le c\right)}_{P_{21}}+\underbrace{\Pr\left(\left|\frac{\widehat{\mathrm{dcov}}^2(\widehat{e}_{j,C},\mathbf{x}_k)}{\widehat{I}_{jk,2}}-\frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C},\mathbf{x}_k)}{\widehat{I}_{jk,2}}\right|\ge\frac{\epsilon}{4}\text{ and }|\widehat{I}_{jk,2}|>c\right)}_{P_{22}}\\
&\quad+\underbrace{\Pr\left(\left|\frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C},\mathbf{x}_k)}{\widetilde{I}_{jk,2}}-\frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C},\mathbf{x}_k)}{\widehat{I}_{jk,2}}\right|\ge\frac{\epsilon}{4}\text{ and }|\widehat{I}_{jk,2}|>c\right)}_{P_{23}}\\
&:= P_{21}+P_{22}+P_{23}.
\end{aligned}$$

Applying Equation (26) and simple calculation, we have

$$\begin{aligned}
P_{21}&\le\Pr\left(\tau_2-|\widetilde{I}_{jk,2}-\widehat{I}_{jk,2}|\le c\right)\le\Pr\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C}-\widetilde{e}_{j,C}\|\ge\frac{1}{4\sqrt{\lambda+1}}(\tau_2-c)\right)+Q_1\exp(-Q_2n+4(|C|+2)).\\
P_{22}&\le\Pr\left(|\widehat{\mathrm{dcov}}^2(\widehat{e}_{j,C},\mathbf{x}_k)-\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C},\mathbf{x}_k)|\ge\frac{\epsilon}{4}c\right).
\end{aligned}$$

Additionally, applying the bound of $|\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C},\mathbf{x}_k)|$ in Equation (16) and Lemma 17,

$$\begin{aligned}
P_{23}&\le\Pr\left(|\widetilde{I}_{jk,2}-\widehat{I}_{jk,2}|\ge\frac{\epsilon}{4}\frac{\widetilde{I}_{jk,2}\widehat{I}_{jk,2}}{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C},\mathbf{x}_k)}\text{ and }|\widehat{I}_{jk,2}|>c\right)\\
&\le\Pr\left(|\widetilde{I}_{jk,2}-\widehat{I}_{jk,2}|\ge\frac{\epsilon}{4}\frac{\tau_2 c}{16(\lambda+1)^{3/2}}\right)+Q_1\exp(-Q_2n+4(|C|+2)).
\end{aligned}$$

Then, Equation (26) yields

$$P_{23}\le\Pr\left(\frac{1}{\sqrt{n}}\|\widehat{e}_{j,C}-\widetilde{e}_{j,C}\|\ge\frac{\epsilon}{4}\frac{\tau_2 c}{64(\lambda+1)^2}\right)+Q_1\exp(-Q_2n+4(|C|+2)).$$

Finally, Lemma 12 and Lemma 11 give the desired results.

$$\Pr\left(|T(e_{j,C},X_k)-\widehat{T}(\hat{e}_{j,C},\mathbf{x}_k)|\ge\epsilon\right)\le P_1+P_{21}+P_{22}+P_{23}=\mathcal{O}\left(\exp\left(-\frac{K_1\epsilon^4}{\lambda^{16}}n+4(|C|+2)\right)\right).$$

□

11.3 Proof for Lemma 10

Proof. The proof proceeds analogously to Lemma 9, with only the distinct points highlighted. Applying the union bound, the error probability can be decomposed as follow.

$$\begin{aligned}
&\Pr\left(|T(e_{j,C},r_C(\ell,j))-\widehat{T}(\hat{e}_{j,C},\hat{r}_C(\ell,j))|\ge\epsilon\right)\\
&\le\underbrace{\Pr\left(|\widehat{T}(\widetilde{e}_{j,C},\widetilde{r}_C(\ell,j))-\widehat{T}(\hat{e}_{j,C},\hat{r}_C(\ell,j))|\ge\frac{\epsilon}{2}\right)}_{P_3}+\underbrace{\Pr\left(|T(e_{j,C},r_C(\ell,j))-\widehat{T}(\widetilde{e}_{j,C},\widetilde{r}_C(\ell,j))|\ge\frac{\epsilon}{2}\right)}_{P_4}:=P_3+P_4.
\end{aligned}$$

Following the proof of Theorem 1 in Li et al. (2012), we have

$$P_4 = \mathcal{O}\left(\exp\left(-\left(\frac{\epsilon^2}{M^2} + Ms\right)n\right)\right),$$

where M is a truncation constant. Notably, none of ϵ, M and s depend on n, p or d_{in}.

Now, we decompose P_3 using a positive constant $c < \tau_3 = \min_{j\neq\ell} \widetilde{J}_{j\ell,2}$ as follows.

$$\begin{aligned} P_3 \leq & \underbrace{\Pr\left(|\widehat{J}_{j\ell,2}| \leq c\right)}_{P_{31}} + \underbrace{\Pr\left(\left|\frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \hat{r}_C(\ell,j))}{\widehat{J}_{j\ell,2}} - \frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))}{\widehat{J}_{j\ell,2}}\right| \geq \frac{\epsilon}{4} \text{ and } |\widehat{J}_{jk,2}| > c\right)}_{P_{32}} \\ & + \underbrace{\Pr\left(\left|\frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))}{\widetilde{J}_{j\ell,2}} - \frac{\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))}{\widehat{J}_{j\ell,2}}\right| \geq \frac{\epsilon}{4} \text{ and } |\widehat{J}_{j\ell,2}| > c\right)}_{P_{33}} \\ := & P_{31} + P_{32} + P_{33}. \end{aligned}$$

Similar to Equation (14), $\max\left\{\widetilde{J}_{j\ell,1}, \widetilde{J}_{j\ell,2}, \widetilde{J}_{j\ell,3}\right\} \leq 4\left(\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C}\|\right)\left(\frac{1}{\sqrt{n}}\|\widetilde{r}_C(\ell,j)\|\right)$, and

$$\frac{1}{\sqrt{n}}\|\widetilde{r}_C(\ell,j)\| \leq \frac{1}{\sqrt{n}}\|\widetilde{e}_{\ell,C}\| + \frac{\Sigma_{\ell,j|C}}{\Sigma_{j|C}}\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C}\| \tag{17}$$

by the triangular inequality. Then, if $\|\widehat{\Sigma}_D - \Sigma_D\|_2 < 1$ with $D = \{j,\ell\} \cup C$, Equation (13) gives that

$$\max\left\{\widetilde{J}_{j\ell,1}, \widetilde{J}_{j\ell,2}, \widetilde{J}_{j\ell,3}\right\} \leq 4(1+\lambda^2)\max\left\{\frac{1}{n}\|\widetilde{e}_{j,C}\|^2, \frac{1}{n}\|\widetilde{e}_{\ell,C}\|\|\widetilde{e}_{j,C}\|\right\} \leq 4(1+\lambda^2)(1+\lambda+\lambda^2).$$

Additionally, applying Lemma 17, there exist some positive constants Q_1 and Q_2 such that

$$\Pr\left(\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j)) \leq 16(1+\lambda^2)(1+\lambda+\lambda^2)\right) \geq 1 - Q_1\exp(-Q_2 n + 4(|C|+2)). \tag{18}$$

Note that

$$\begin{aligned} |\widetilde{J}_{j\ell,2} - \widehat{J}_{j\ell,2}| = & \left|\left(\frac{1}{n^2}\sum_{i,h=1}^n |\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(h)}|\right)\left(\frac{1}{n^2}\sum_{i,h=1}^n |\widetilde{r}_C(\ell,j)^{(i)} - \widetilde{r}_C(\ell,j)^{(h)}| - |\hat{r}_C(\ell,j)^{(i)} - \hat{r}_C(\ell,j)^{(h)}|\right)\right. \\ & \left. + \left(\frac{1}{n^2}\sum_{i,h=1}^n |\widetilde{r}_C(\ell,j)^{(i)} - \widetilde{r}_C(\ell,j)^{(h)}|\right)\left(\frac{1}{n^2}\sum_{i,h=1}^n |\widetilde{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(h)}| - |\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(h)}|\right)\right| \\ \leq & \left(\frac{2}{n}\sum_{i=1}^n |\hat{e}_{j,C}^{(i)}|\right)\left(\frac{2}{n}\sum_{i=1}^n |\widetilde{r}_C(\ell,j)^{(i)} - \hat{r}_C(\ell,j)^{(i)}|\right) + \left(\frac{2}{n}\sum_{i=1}^n |\widetilde{r}_C(\ell,j)^{(i)}|\right)\left(\frac{2}{n}\sum_{i=1}^n |\widetilde{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(i)}|\right) \\ \leq & 4\frac{1}{\sqrt{n}}\|\hat{e}_{j,C}\|\frac{1}{\sqrt{n}}\|\widetilde{r}_C(\ell,j) - \hat{r}_C(\ell,j)\| + 4\frac{1}{\sqrt{n}}\|\widetilde{r}_C(\ell,j)\|\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C} - \hat{e}_{j,C}\|. \end{aligned}$$

By applying Equation (17) and triangular inequality,

$$\begin{aligned} |\widetilde{J}_{j\ell,2} - \widehat{J}_{j\ell,2}| \leq & 4\left(\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C}\| + \frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C} - \hat{e}_{j,C}\|\right)\frac{1}{\sqrt{n}}\|\widetilde{r}_C(\ell,j) - \hat{r}_C(\ell,j)\| \\ & + 4\left(\frac{1}{\sqrt{n}}\|\widetilde{e}_{\ell,C}\| + \lambda^2\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C}\|\right)\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C} - \hat{e}_{j,C}\|. \end{aligned}$$

Let $D = \{j, \ell\} \cup C$. Suppose that $\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 < 1$. Applying Equations (13),

$$|\widetilde{J}_{j\ell,2} - \widehat{J}_{j\ell,2}| \leq \underbrace{4(\sqrt{1+\lambda+\lambda^2} + \frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C} - \hat{e}_{j,C}\|)\frac{1}{\sqrt{n}}\|\widetilde{r}_C(\ell,j) - \hat{r}_C(\ell,j)\|}_{J_1} + \underbrace{4(1+\lambda^2)\sqrt{1+\lambda+\lambda^2}\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C} - \hat{e}_{j,C}\|}_{J_2}. \tag{19}$$

Given Equation (19), applying Lemma 11,

$$\Pr\left(J_2 > \frac{\tau_3 - c}{2}\right) \leq \mathcal{O}\left(\exp\left(-\frac{M_2(\tau_3 - c)^4}{\lambda^{20}}n + 4(|C|+2)\right)\right). \tag{20}$$

Subsequently,

$$\begin{aligned}\Pr\left(J_1 > \frac{\tau_3 - c}{2}\right) &\leq \Pr\left(J_2 > \frac{\tau_3 - c}{2}\right) + \Pr\left(J_2 \leq \frac{\tau_3 - c}{2} \text{ and } J_1 > \frac{\tau_3 - c}{2}\right) \\ &\leq \Pr\left(J_2 > \frac{\tau_3 - c}{2}\right) + \Pr\left(\frac{1}{\sqrt{n}}\|\widetilde{r}_C(\ell,j) - \hat{r}_C(\ell,j)\| > \frac{\tau_3 - c}{8\sqrt{1+\lambda+\lambda^2} + \tau_3 - c}\right)\end{aligned}$$

Then, Equation (20) and Lemma 13 give

$$\begin{aligned}\Pr\left(J_1 > \frac{\tau_3 - c}{2}\right) &\leq \mathcal{O}\left(\exp\left(-\frac{M_2(\tau_3 - c)^4}{\lambda^{20}}n + 4(|C|+2)\right)\right) + \mathcal{O}\left(\exp\left(-\frac{M_2(\tau_3 - c)^4}{\lambda^{16}}n + 4(|C|+2)\right)\right) \\ &\leq \mathcal{O}\left(\exp\left(-\frac{M_2(\tau_3 - c)^4}{\lambda^{20}}n + 4(|C|+2)\right)\right).\end{aligned} \tag{21}$$

Combining Equation (19), Equation (20), and Equation (21) with Lemma 17,

$$P_{31} \leq \Pr\left(|\widetilde{J}_{j\ell,2} - \widehat{J}_{j\ell,2}| > \tau_3 - c\right) \leq \mathcal{O}\left(\exp\left(-\frac{M_2(\tau_3 - c)^4}{\lambda^{20}}n + 4(|C|+2)\right)\right).$$

Additionally, Lemma 14 gives

$$P_{32} \leq \Pr\left(|\widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \hat{r}_C(\ell,j)) - \widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))| \geq \frac{\epsilon}{4}c\right) \leq \mathcal{O}\left(\exp\left(-\frac{M_2 c^4 \epsilon^4}{\lambda^{16}}n + 4(|C|+2)\right)\right).$$

Furthermore, applying the bound of $\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))$ in Equation (18) and Lemma 17,

$$P_{33} \leq \Pr\left(|\widetilde{J}_{j\ell,2} - \widehat{J}_{j\ell,2}| \geq \frac{\epsilon}{4}\frac{\tau_3 c}{16(1+\lambda^2)(1+\lambda+\lambda^2)}\right) + Q_1\exp(-Q_2 n + 4(|C|+2)).$$

Then, using a similar argument as in P_{31},

$$P_{33} \leq \mathcal{O}\left(\exp\left(-\frac{M_2(\tau_3 c)^4}{\lambda^{36}}n + 4(|C|+2)\right)\right).$$

Finally, combining all of the results, there exists a positive constant K_2 such that

$$\Pr\left(|T(e_{j,C}, r_C(\ell,j)) - \widehat{T}(\hat{e}_{j,C}, \hat{r}_C(\ell,j))| \geq \epsilon\right) \leq \mathcal{O}\left(\exp\left(-\frac{K_2\epsilon^4}{\lambda^{36}}n + 4(|C|+2)\right)\right).$$

□

12 Useful Lemmas

Lemma 11. *Suppose that Assumption 3 is satisfied. For any $j \in V$, $C \subset V \setminus \{j\}$ and $\epsilon > 0$, there exists a constant M_1 such that*

$$\Pr\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\| \geq \epsilon\right) \leq exp\left(-\frac{M_1\epsilon^4}{(\lambda^2(2\lambda^2+2\lambda+4)+\lambda\epsilon^2)^2}n + 4(|C|+1)\right).$$

Proof. Let $D = \{j\} \cup C$, then the covariance matrix of (X_j, X_C) is

$$\Sigma_{D,D} = \begin{bmatrix} \Sigma_{j,j} & \Sigma_{j,C} \\ \Sigma_{C,j} & \Sigma_{C,C} \end{bmatrix}.$$

Under Assumption 3, the min-max principle for singular values derives

$$\max\{\|\Sigma_{j,C}\|, \|\Sigma_{C,C}\|_2\} \leq \|\Sigma_{D,D}\|_2 \leq \|\Sigma\|_2 \leq \lambda. \tag{22}$$

Note that

$$\begin{aligned}\frac{1}{n}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\Sigma}_{j,C}(\widehat{\Sigma}_{C,C})^{-1} - \Sigma_{j,C}(\Sigma_{C,C})^{-1}\right)\mathbf{x}_C^{(i)}\mathbf{x}_C^{(i)T}\left((\widehat{\Sigma}_{C,C})^{-1}\widehat{\Sigma}_{C,j} - (\Sigma_{C,C})^{-1}\Sigma_{C,j}\right) \\ &= \left(\widehat{\Sigma}_{j,C}(\widehat{\Sigma}_{C,C})^{-1} - \Sigma_{j,C}(\Sigma_{C,C})^{-1}\right)\widehat{\Sigma}_{C,C}\left((\widehat{\Sigma}_{C,C})^{-1}\widehat{\Sigma}_{C,j} - (\Sigma_{C,C})^{-1}\Sigma_{C,j}\right).\end{aligned}$$

Hence, by using the sub-multiplicativity of a matrix norm and the inequality in Equation (22),

$$\begin{aligned}\frac{1}{n}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|^2 &= \widehat{\Sigma}_{j,C}[(\widehat{\Sigma}_{C,C})^{-1} - (\Sigma_{C,C})^{-1}]\widehat{\Sigma}_{C,j} + [\widehat{\Sigma}_{j,C} - \Sigma_{j,C}](\Sigma_{C,C})^{-1}\widehat{\Sigma}_{C,j} \\ &+ \Sigma_{j,C}(\Sigma_{C,C})^{-1}[\widehat{\Sigma}_{C,C} - \Sigma_{C,C}](\Sigma_{C,C})^{-1}\Sigma_{C,j} + [\Sigma_{j,C} - \widehat{\Sigma}_{j,C}](\Sigma_{C,C})^{-1}\Sigma_{Cj} \\ &\leq \underbrace{\|(\widehat{\Sigma}_{C,C})^{-1} - (\Sigma_{C,C})^{-1}\|_2}_{A_1}\|\widehat{\Sigma}_{C,j}\|^2 + \lambda\|\widehat{\Sigma}_{j,C} - \Sigma_{j,C}\|\underbrace{\|\widehat{\Sigma}_{C,j}\|}_{A_2} + \lambda^4\underbrace{\|\widehat{\Sigma}_{C,C} - \Sigma_{C,C}\|_2}_{A_3} + \lambda^2\underbrace{\|\widehat{\Sigma}_{j,C} - \Sigma_{jC}\|}_{A_4}.\end{aligned}$$

Suppose that $\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 < \frac{\delta}{\lambda^2+\lambda\delta} < \frac{1}{\lambda}$ for any positive constant δ. Then, applying Equation (22) and Lemma 15, we have

$$\begin{aligned}&A_1 < \delta, \quad A_2 \leq \|\widehat{\Sigma}_{j,C} - \Sigma_{j,C}\| + \|\Sigma_{j,C}\| \leq \|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 + \lambda < \lambda + 1, \\ &\max\{A_3,\, A_4\} \leq \|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 < \frac{\delta}{\lambda^2+\lambda\delta}.\end{aligned}$$

Hence, for any positive constant $\epsilon > 0$, if $\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 < \frac{\delta}{\lambda^2+\lambda\delta}$ with $\delta = \frac{\epsilon^2}{2\lambda^2+2\lambda+4}$,

$$\frac{1}{n}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|^2 \leq (\lambda^4+2\lambda^2+\lambda)\frac{\delta}{\lambda^2+\lambda\delta} + (\lambda+1)^2\delta \leq \epsilon^2. \tag{23}$$

Consequently, using Lemma 17, there exists a positive constant M_1 such that

$$\begin{aligned}\Pr\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\| \geq \epsilon\right) &\leq \Pr\left(\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 \geq \frac{\epsilon^2}{\lambda^2(2\lambda^2+2\lambda+4)+\lambda\epsilon^2}\right) \\ &\leq \exp\left(-\frac{M_1\epsilon^4}{(\lambda^2(2\lambda^2+2\lambda+4)+\lambda\epsilon^2)^2}n + 4(|C|+1)\right).\end{aligned}$$

□

Lemma 12. *Suppose that Assumption 3 is satisfied. Consider the sample and estimated versions of* $\widehat{dcov}^2$, *as defined in Equations (7) and (8). For any* $j \in V$, *sets* $S \subset V \setminus \{j\}$ *and* $C \subset V \setminus \{j\}$, *and* $\epsilon > 0$, *there exists a constant* M_1 *such that*

$$\Pr\left(|\widehat{dcov}^2(\widehat{e}_{j,C}, \mathbf{x}_S) - \widehat{dcov}^2(\widetilde{e}_{j,C}, \mathbf{x}_S)| \geq \epsilon\right) = \mathcal{O}\left(exp\left(-\frac{M_1\epsilon^4}{\lambda^{10}|S|^2}n + 4(|C| + |S| + 1)\right)\right).$$

Proof. This proof is similar as that of Lemma 5 in Zhao et al. (2022), while the result is more general than that of Lemma 5 in Zhao et al. (2022). Applying Equations (7) and (8),

$$|\widehat{\text{dcov}}^2(\widetilde{e}_{j,C}, \mathbf{x}_S) - \widehat{\text{dcov}}^2(\hat{e}_{j,C}, \mathbf{x}_S)| \leq |\widehat{I}_{jS,1} - \widetilde{I}_{jS,1}| + |\widehat{I}_{jS,2} - \widetilde{I}_{jS,2}| + 2|\widehat{I}_{jS,3} - \widetilde{I}_{jS,3}|. \tag{24}$$

Simple algebra yields that

$$\begin{aligned}
|\widetilde{I}_{jS,1} - \widehat{I}_{jS,1}| &\leq \frac{1}{n^2}\sum_{i,h=1}^{n} \|\mathbf{x}_S^{(i)} - \mathbf{x}_S^{(h)}\| \left||\hat{e}_{j,C}^{(i)} - \hat{e}_{j,C}^{(h)}| - |\widetilde{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(h)}|\right| \leq \frac{2}{n^2}\sum_{i,h=1}^{n} \|\mathbf{x}_S^{(i)} - \mathbf{x}_S^{(h)}\||\hat{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(i)}| \\
&\leq \frac{2}{n}\sum_{i=1}^{n} \|\mathbf{x}_S^{(i)}\||\hat{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(i)}| + 2\left(\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_S^{(i)}\|\right)\left(\frac{1}{n}\sum_{i=1}^{n}|\hat{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(i)}|\right).
\end{aligned}$$

By Cauchy-Schwarz inequality and Jensen's inequality, we have

$$\begin{aligned}
|\widetilde{I}_{jS,1} - \widehat{I}_{jS,1}| &\leq 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_S^{(i)T}\mathbf{x}_S^{(i)}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(i)}|^2} + 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_S^{(i)T}\mathbf{x}_S^{(i)}}\left(\frac{1}{n}\sum_{i=1}^{n}|\hat{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(i)}|\right) \\
&\leq 4\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_S^{(i)T}\mathbf{x}_S^{(i)}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(i)}|^2}.
\end{aligned}$$

Additionally, by the arithmetic mean and quadratic mean inequality, we have

$$|\widetilde{I}_{jS,1} - \widehat{I}_{jS,1}| \leq 4\sqrt{tr(\widehat{\Sigma}_{S,S})}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{e}_{j,C}^{(i)} - \widetilde{e}_{j,C}^{(i)}|^2}.$$

Then, we can use an analogous approach in the proof for Lemma 9. Let $D = \{j\} \cup C \cup S$. Then, if $\|\widehat{\Sigma}_D - \Sigma_D\|_2 < 1$, Assumption 3 gives that

$$|\widetilde{I}_{jS,1} - \widehat{I}_{jS,1}| \leq 4\sqrt{(\lambda+1)|S|}\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|. \tag{25}$$

Additionally, following a similar calculation, we also have

$$\begin{aligned}
|\widetilde{I}_{jS,2} - \widehat{I}_{jS,2}| &\leq 4\sqrt{(\lambda+1)|S|}\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|, \\
|\widetilde{I}_{jS,3} - \widehat{I}_{jS,3}| &\leq 4\sqrt{(\lambda+1)|S|}\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|.
\end{aligned} \tag{26}$$

By plugging Equations (25) and (26) into Equation (24), we have

$$|\widehat{\text{dcov}}^2(\widetilde{e}_{j,C}, \mathbf{x}_S) - \widehat{\text{dcov}}^2(\hat{e}_{j,C}, \mathbf{x}_S)| \leq 16\sqrt{(\lambda+1)|S|}\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|. \tag{27}$$

Applying Lemma 11 and Lemma 17, there exist positive constants M_1, Q_1 and Q_2 such that

$$\begin{aligned}
&\Pr\left(|\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \mathbf{x}_S) - \widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \mathbf{x}_S)| \geq \epsilon\right) \\
&\quad \leq \Pr\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\| \geq \frac{\epsilon}{16\sqrt{(\lambda+1)|S|}}\right) + Q_1 \exp(-Q_2 n + 4(|C| + |S| + 1)) \\
&\quad \leq \exp\left(-\frac{M_1 \epsilon^4}{(16^2|S|\lambda^2(\lambda+1)(2\lambda^2+2\lambda+4) + \lambda\epsilon^2)^2} n + 4(|C|+1)\right) + Q_1 \exp(-Q_2 n + 4(|C| + |S| + 1)).
\end{aligned}$$

Finally, the desired result is obtained as follows.

$$\Pr\left(|\widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \mathbf{x}_S) - \widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \mathbf{x}_S)| \geq \epsilon\right) = \mathcal{O}\left(\exp\left(-\frac{M_1 \epsilon^4}{\lambda^{10}|S|^2} n + 4(|C| + |S| + 1)\right)\right).$$

□

Lemma 13. *Suppose that Assumption 3 is satisfied. For any $j \in V$, $C \subset V \setminus \{j\}$ and $\epsilon > 0$, there exists a constant M_2 such that*

$$\Pr\left(\frac{1}{\sqrt{n}}\|\hat{r}_C(j,k) - \widetilde{r}_C(j,k)\| \geq \epsilon\right) \leq \exp\left(-\frac{M_2 \epsilon^4}{(\lambda(3\lambda^5 + 6\lambda^4 + 15\lambda^3 + 10\lambda^2 + 9\lambda) + \lambda\epsilon^2)^2} n + 4(|C| + 2)\right).$$

Proof. The idea of proof is similar to the proof for Lemma 11. First, note that

$$\begin{aligned}
\hat{r}_C(j,k) - \widetilde{r}_C(j,k) &= (\hat{e}_{j,C} - \widetilde{e}_{j,C}) + \frac{\Sigma_{j,k|C}}{\Sigma_{k|C}}(\widetilde{e}_{k,C} - \hat{e}_{k,C}) + \left(\frac{\Sigma_{j,k|C}}{\Sigma_{k|C}} - \frac{\widehat{\Sigma}_{j,k|C}}{\widehat{\Sigma}_{k|C}}\right)\hat{e}_{k,C} \\
&= (\hat{e}_{j,C} - \widetilde{e}_{j,C}) + \frac{\Sigma_{j,k|C}}{\Sigma_{k|C}}(\widetilde{e}_{k,C} - \hat{e}_{k,C}) - \frac{1}{\widehat{\Sigma}_{k|C}}\left(\widehat{\Sigma}_{j,k|C} - \Sigma_{j,k|C}\right)\hat{e}_{k,C} - \Sigma_{j,k|C}\left(\frac{1}{\widehat{\Sigma}_{k|C}} - \frac{1}{\Sigma_{k|C}}\right)\hat{e}_{k,C}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
\frac{1}{n}\|\hat{r}_C(j,k) - \widetilde{r}_C(j,k)\|^2 \leq \frac{1}{n}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|^2 + \frac{\lambda^2}{n}\|\hat{e}_{k,C} - \widetilde{e}_{k,C}\|^2 + \left|\frac{1}{\widehat{\Sigma}_{k|C}}\left(\widehat{\Sigma}_{j,k|C} - \Sigma_{j,k|C}\right)\right| \cdot \frac{1}{n}\|\hat{e}_{k,C}\|^2 \\
+ \left|\Sigma_{j,k|C}\left(\frac{1}{\widehat{\Sigma}_{k|C}} - \frac{1}{\Sigma_{k|C}}\right)\right| \cdot \frac{1}{n}\|\hat{e}_{k,C}\|^2.
\end{aligned}$$

Let $D = \{j,k\} \cup C$. Suppose that $\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 < \frac{\delta}{\lambda^2+\lambda\delta} < \frac{1}{\lambda}$ for any positive constant δ. Then, applying Equation (22) along with an analogous computation in Lemma 15, we have

$$\begin{aligned}
\frac{1}{n}\|\hat{r}_C(j,k) - \widetilde{r}_C(j,k)\|^2 &\leq \frac{1}{n}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|^2 + \lambda^2\frac{1}{n}\|\hat{e}_{k,C} - \widetilde{e}_{k,C}\|^2 \\
&\quad + \left(\frac{(2\lambda^2+\lambda+1)\delta}{\lambda^2+\lambda\delta} + (\lambda+1)^2\delta + \lambda\delta\right)\left(\frac{1}{n}\|\hat{e}_{k,C} - \widetilde{e}_{k,C}\|^2 + \frac{1}{n}\|\widetilde{e}_{k,C}\|^2\right) \\
&\leq \left(1 + \lambda^2 + \frac{(2\lambda^2+\lambda+1)\delta}{\lambda^2+\lambda\delta} + (\lambda+1)^2\delta + \lambda\delta\right)\max\left\{\frac{1}{n}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|^2, \frac{1}{n}\|\hat{e}_{k,C} - \widetilde{e}_{k,C}\|^2\right\} \\
&\quad + \left(\frac{(2\lambda^2+\lambda+1)\delta}{\lambda^2+\lambda\delta} + (\lambda+1)^2\delta + \lambda\delta\right)\frac{1}{n}\|\widetilde{e}_{j,C}\|^2.
\end{aligned}$$

Subsequently, applying Equation (23) and Equation (13),

$$\begin{aligned}
\frac{1}{n}\|\hat{r}_C(j,k) - \widetilde{r}_C(j,k)\|^2 &\leq \left(1 + \lambda^2 + \frac{(2\lambda^2+\lambda+1)\delta}{\lambda^2+\lambda\delta} + (\lambda+1)^2\delta + \lambda\delta\right)\left(\frac{(\lambda^4+2\lambda^2+\lambda)\delta}{\lambda^2+\lambda\delta} + (\lambda+1)^2\delta\right) \\
&\quad + \left(\frac{(2\lambda^2+\lambda+1)\delta}{\lambda^2+\lambda\delta} + (\lambda+1)^2\delta + \lambda\delta\right)(1 + \lambda + \lambda^2).
\end{aligned}$$

Simplifying the form using the fact that $\frac{\delta}{\lambda^2+\lambda\delta} < \frac{\delta}{\lambda^2}$ and $\frac{1}{\lambda}, \frac{1}{\lambda^2} < 1$,

$$\frac{1}{n}\|\hat{r}_C(j,k) - \widetilde{r}_C(j,k)\|^2 \leq \left(1+\lambda^2+(\lambda^2+3\lambda+5)\delta\right)(2\lambda^2+2\lambda+4)\delta + (\lambda^2+3\lambda+5)(\lambda^2+\lambda+1)\delta \tag{28}$$

Hence, for a small $\epsilon > 0$, if $\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 < \frac{\delta}{\lambda^2+\lambda\delta} < \frac{1}{\lambda}$ with $\delta = \delta_3 := \frac{\epsilon^2}{3\lambda^4+6\lambda^3+15\lambda^2+10\lambda+9}$,

$$\frac{1}{\sqrt{n}}\|\hat{r}_C(j,k) - \widetilde{r}_C(j,k)\| \leq \epsilon.$$

Consequently, using Lemma 17, there exists a positive constant M_2 such that

$$\begin{aligned}\Pr\left(\frac{1}{\sqrt{n}}\|\hat{r}_C(j,k) - \widetilde{r}_C(j,k)\| \geq \epsilon\right) &\leq \Pr\left(\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 \geq \frac{\delta_3}{\lambda^2+\lambda\delta_3}\right) \\ &\leq \exp\left(-\frac{M_2\epsilon^4}{(\lambda(3\lambda^5+6\lambda^4+15\lambda^3+10\lambda^2+9\lambda)+\lambda\epsilon^2)^2}n + 4(|C|+2)\right).\end{aligned}$$

□

Lemma 14. *Suppose that Assumption 3 is satisfied. Consider the sample and estimated versions of* $\widehat{dcov}^2$, *as defined in Equations* (10) *and* (11). *For any* $j, \ell \in V$, *a set* $C \subset V \setminus \{j,\ell\}$ *and* $\epsilon > 0$, *there exists a constant* M_2 *such that*

$$\Pr\left(|\widehat{dcov}^2(\hat{e}_{j,C}, \hat{r}_C(\ell,j)) - \widehat{dcov}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))| \geq \epsilon\right) = \mathcal{O}\left(exp\left(-\frac{M_2\epsilon^4}{\lambda^{16}}n + 4(|C|+2)\right)\right),$$

where $\widetilde{r}_C(j,k) = \widetilde{e}_{j,C} - \frac{Cov(e_{j,C}, e_{k,C})}{Var(e_{k,C})}\widetilde{e}_{k,C}$.

Proof. The proof of Lemma 14 is similar as that of Lemma 12. First, we decompose the term as follows.

$$\begin{aligned}&\left|\widehat{\text{dcov}}^2(\hat{e}_{j,C}, \hat{r}_C(\ell,j)) - \widehat{\text{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))\right| \\ &\leq \underbrace{\left|\widehat{\text{dcov}}^2(\hat{e}_{j,C}, \hat{r}_C(\ell,j) - \widehat{\text{dcov}}^2(\widetilde{e}_{j,C}, \hat{r}_C((\ell,j))\right|}_{I_1} + \underbrace{\left|\widehat{\text{dcov}}^2(\widetilde{e}_{j,C}, \hat{r}_C(\ell,j)) - \widehat{\text{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))\right|}_{I_2} := I_1 + I_2.\end{aligned}$$

Applying Equation (27) and appropriately replacing $(\lambda+1)|S|$, which is derived from $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_S^{(i)T}\mathbf{x}_S^{(i)}$, I_1 and I_2 are also bounded as follows.

$$\begin{aligned}I_1 &\leq 16\sqrt{\frac{1}{n}\sum_{i=1}^n \hat{r}_C^{(i)T}(\ell,j)\hat{r}_C^{(i)}(\ell,j)}\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|\right) \leq 16\sqrt{\widehat{\Sigma}_{\ell|C} - \widehat{\Sigma}_{\ell,j|C}(\widehat{\Sigma}_{j|C})^{-1}\widehat{\Sigma}_{j,\ell|C}}\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|\right), \\ I_2 &\leq 16\left(\frac{1}{\sqrt{n}}\|\widetilde{e}_{j,C}\|\right)\left(\frac{1}{\sqrt{n}}\|\hat{r}_C(\ell,j) - \widetilde{r}_C(\ell,j)\|\right).\end{aligned}$$

Then, we can use analogous an approach in the proof for Lemma 9. Let $D = \{j,\ell\} \cup C$. If $\|\widehat{\Sigma}_D - \Sigma_D\|_2 < 1$, applying Equation (13),

$$\begin{aligned}I_1 &\leq 16\sqrt{\widehat{\Sigma}_{\ell|C}}\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|\right) \leq 16\frac{1}{\sqrt{n}}\|\hat{e}_{\ell,C}\|\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|\right) \\ &\leq 16\left(\sqrt{\lambda^2+\lambda+1} + \frac{1}{\sqrt{n}}\|\hat{e}_{\ell,C} - \widetilde{e}_{\ell,C}\|\right)\left(\frac{1}{\sqrt{n}}\|\hat{e}_{j,C} - \widetilde{e}_{j,C}\|\right).\end{aligned}$$

Additionally, if $\|\widehat{\Sigma}_D - \Sigma_D\|_2 < 1$, by Assumption 3 and Equation (13),

$$I_2 \leq 16\sqrt{\lambda^2+\lambda+1}\left(\frac{1}{\sqrt{n}}\|\hat{r}_C(\ell,j) - \widetilde{r}_C(\ell,j)\|\right).$$

For a small ϵ, applying Lemma 11, Lemma 13 and Lemma 17, we have

$$\Pr\left(I_1 \geq \frac{\epsilon}{2}\right) \leq \mathcal{O}\left(\exp\left(-\frac{M_2\epsilon^4}{\lambda^{12}}n + 4(|C|+2)\right)\right),$$
$$\Pr\left(I_2 \geq \frac{\epsilon}{2}\right) \leq \mathcal{O}\left(\exp\left(-\frac{M_2\epsilon^4}{\lambda^{16}}n + 4(|C|+2)\right)\right).$$

Consequently, applying the union bound, there exists a positive constant M_2 such that

$$\begin{aligned}\Pr\left(|\widehat{\mathrm{dcov}}^2(\hat{e}_{j,C}, \hat{r}_C(\ell,j)) - \widehat{\mathrm{dcov}}^2(\widetilde{e}_{j,C}, \widetilde{r}_C(\ell,j))| \geq \epsilon\right) &\leq \Pr\left(I_1 \geq \frac{\epsilon}{2}\right) + \Pr\left(I_2 \geq \frac{\epsilon}{2}\right) \\ &= \mathcal{O}\left(\exp\left(-\frac{M_2\epsilon^4}{\lambda^{16}}n + 4(|C|+2)\right)\right).\end{aligned}$$

□

Lemma 15. *Consider a sub-Gaussian LiNGAM under Assumption 3. In addition, Σ is the true covariance matrix and $\Sigma_{A,B}$ is the $|A| \times |B|$ sub-matrix of Σ corresponding to random vectors X_A and X_B. Then for any $j \in V$, a small $C \subset V \setminus \{j\}$, and a positive constant $\delta > 0$,*

$$\text{if} \quad \|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 \leq \frac{\delta}{\lambda^2 + \lambda\delta} \quad \text{, then} \quad \|(\widehat{\Sigma}_{C,C})^{-1} - (\Sigma_{C,C})^{-1}\|_2 \leq \delta,$$

where $D = \{j\} \cup C$.

Proof. Under Assumption 3 and the min-max principle for singular values,

$$\|\Sigma_{C,C}\|_2 \leq \|\Sigma_{D,D}\|_2 \leq \|\Sigma\|_2 \leq \lambda. \tag{29}$$

By applying Equation (29), sub-multiplicativity of the induced ℓ_2-matrix norm and the triangular inequality yield

$$\begin{aligned}\|(\widehat{\Sigma}_{C,C})^{-1} - (\Sigma_{C,C})^{-1}\|_2 &\leq \|(\widehat{\Sigma}_{C,C})^{-1}\|_2\|\widehat{\Sigma}_{C,C} - \Sigma_{C,C}\|_2\|(\Sigma_{C,C})^{-1}\|_2 \\ &\leq \lambda\|(\widehat{\Sigma}_{C,C})^{-1} - (\Sigma_{C,C})^{-1}\|_2\|\widehat{\Sigma}_{C,C} - \Sigma_{C,C}\|_2 + \lambda^2\|\widehat{\Sigma}_{C,C} - \Sigma_{C,C}\|_2 \\ &\leq \lambda\|(\widehat{\Sigma}_{C,C})^{-1} - (\Sigma_{C,C})^{-1}\|_2\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 + \lambda^2\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2.\end{aligned}$$

Hence, if $\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2 \leq \frac{\delta}{\lambda^2+\lambda\delta}$, then

$$\|[\widehat{\Sigma}_{C,C}]^{-1} - [\Sigma_{C,C}]^{-1}\|_2 \leq \frac{\lambda^2\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2}{1 - \lambda\|\widehat{\Sigma}_{D,D} - \Sigma_{D,D}\|_2} \leq \delta.$$

□

Lemma 16 (Theorem 6.5 (a) in Wainwright, 2019)**.** *For any row-wise σ-sub Gaussian random matrix $\mathbf{x}^{1:n} \in \mathbb{R}^{n\times p}$, there exists a positive constant Q_0 such that the sample covariance $\widehat{\Sigma}$ satisfies the bounds*

$$\mathbb{E}\left[exp\left(t\|\widehat{\Sigma} - \Sigma\|_2\right)\right] \leq exp\left(Q_0\frac{t^2\sigma^4}{n} + 4p\right),$$

for all $|t| < \frac{n}{64e^2\sigma^2}$.

Since Lemma 16 and its proof are the same as Theorem 6.5 (a) in Wainwright (2019), we omit the proof. By using Lemma 16 and Chernoff bound, the probability bound for the error of the sample covariance matrix was obtained in Wainwright (2019). This result is slightly modified to make it convenient to use, and we obtain the following lemma.

Lemma 17. *For any row-wise σ-sub Gaussian random matrix $\mathbf{x}^{1:n} \in \mathbb{R}^{n\times p}$, there are positive constants Q_1 and Q_2 such that the sample covariance $\widehat{\Sigma}$ satisfies the probability bound*

$$\Pr\left(\|\widehat{\Sigma} - \Sigma\|_2 \geq \delta\right) \leq Q_1 exp(-Q_2 n \min\{\delta, \delta^2\}),$$

for all $\delta > 0$.

Proof. By applying Chernoff bound and Lemma 16 to the probability of the sample covariance error bound,

$$\Pr\left(\|\widehat{\Sigma} - \Sigma\|_2 \geq \delta\right) \leq \inf_{0\leq|t|\leq n/64e^2\sigma^2} \frac{\mathbb{E}\left[\exp\left(t\|\widehat{\Sigma} - \Sigma\|_2\right)\right]}{\exp(t\delta)} \leq \inf_{0\leq|t|\leq n/64e^2\sigma^2} \frac{\exp(Q_0 t^2\sigma^4/n + 4p)}{\exp(t\delta)}.$$

Then, by introducing $s = \sigma^2 t$, the following result is obtained.

$$\Pr\left(\|\widehat{\Sigma} - \Sigma\|_2 \geq \delta\right) \leq \inf_{0\leq|s|\leq n/64e^2} \exp\left(\frac{Q_0 s^2}{n} + 4p - \frac{\delta s}{\sigma^2}\right) = \begin{cases} \exp\left(-\frac{\delta^2 n}{4Q_0\sigma^4} + 4p\right) & \text{if } \delta \leq \frac{Q_0\sigma^2}{32e^2}, \\ \exp\left(-\frac{\delta n}{64e^2\sigma^2} + \frac{nQ_0}{(64e^2)^2} + 4p\right) & \text{otherwise.} \end{cases}$$

□

13 Proof for Lower Bound of Sample Complexity

Let $\mathcal{G}_{p,d_{in}}$ be the class of all DAGs with p nodes and the maximum indegree d_{in}. Furthermore, $\mathcal{M}_{p,d_{in}}(\lambda)$ denotes the class of p-node sub-Gaussian LiNGAMs where the maximum indegree is $d_{in} \leq \frac{p}{2}$, and Assumption 3 holds in which the eigenvalues of the covariance matrix are in $[\frac{1}{\lambda}, \lambda]$. Lastly, $\mathrm{KL}(M_a \parallel M_b)$ denotes a Kullback-Leibler (KL) divergence between models M_a and M_b, based on M_a.

Lemma 6. *Suppose that $G(M)$ denotes a true DAG corresponding to a model $M \in \mathcal{M}_{p,d_{in}}(\lambda)$. For any positive constant $\delta \in (0, \frac{1}{2})$, there are some positive constants K_1 and K_2 such that if the sample size is*

$$n \leq (1-2\delta)K_1 \frac{d_{in}\log(p/d_{in})}{\log(\lambda)},$$

then, for any estimator $\widehat{G}$,

$$\sup_{M\in\mathcal{M}_{p,d_{in}}(\lambda)} \Pr(\widehat{G} \neq G(M)) \geq \delta - \frac{K_2}{pd_{in}\log(p/d_{in})}.$$

Proof. This proof is analogous to the proof of Theorem 3.1 in Gao et al. (2022) where the lower bound of a sample size for learning a Gaussian linear SEM with the same error variances is discussed. However, our result consider sub-Gaussian distributions, excluding Gaussian, and imposes no restrictions on error variances.

We first compute an upper bound of the KL divergence between any two models in appropriate subset of $\mathcal{M}_{p,d_{in}}(\lambda)$ and the cardinality of $\mathcal{G}_{p,d_{in}}$, denoted as N. In the case of the cardinality, Gao et al. (2022) show that $\log N = \Theta(pd_{in}\log\frac{p}{d_{in}})$ from Lemma B.3 of itself when $d_{in} \leq p/2$.

In the case of KL divergence, we focus on a set of Uniform linear SEMs where KL divergence between each pair of models is well-defined, which is an obvious subset of $\mathcal{M}_{p,d_{in}}(\lambda)$. Let M_s and M_t be such Uniform linear SEMs in $\mathcal{M}_{p,d_{in}}(\lambda)$. The random vectors $X_s \sim M_s$ and $X_t \sim M_t$ can be represented as $X_s = A_s U_s$ and $X_t = A_t U_t$, where U_s and U_t are independent uniform errors. Here $A_s = (I - B_s)^{-1}$ and $A_t = (I - B_t)^{-1}$ where B_s and B_t are edge weight matrices, can be permuted by applying the same row and column permutations simultaneously to transform them into lower triangular matrices with diagonal entries equal to 1, based on the acyclicity assumption.

Combining the fact that $|A_s| = |A_t| = 1$ and the change of variables in probability density function, we have

$$\mathrm{KL}(M_s \parallel M_t) = \mathrm{KL}(U_s \parallel U_t) = \sum_{j=1}^{p} \log\left(\frac{b_{tj} - a_{tj}}{b_{sj} - a_{sj}}\right),$$

where $u_{tj} \in U_t$ and $u_{sj} \in U_s$ are from $U(a_{tj}, b_{tj})$ and $U(a_{sj}, b_{sj})$, respectively, for $j = 1, ..., p$.

Since $\mathrm{Var}(U) = \frac{1}{12}(b-a)^2$ for $U \sim U(a, b)$, both $\frac{1}{12}(b_{tj} - a_{tj})^2$ and $\frac{1}{12}(b_{sj} - a_{sj})^2$ are in $[\frac{1}{\lambda}, \lambda]$ under Assumption 3. Hence, we have

$$\mathrm{KL}(M_s \parallel M_t) \leq p \times \max_{j \in V} \log\left(\frac{b_{tj} - a_{tj}}{b_{sj} - a_{sj}}\right) \leq p \log(\lambda). \tag{30}$$

Applying Equation (30) and $\log N = \Theta(p d_{in} \log \frac{p}{d_{in}})$ to Corollary B.2, we can establish the lower bound of the sample complexity. Consequently, for any positive constant $\delta \in (0, 1/2)$, there are some positive constants $K_1 > 0$ and $K_2 > 0$ such that

$$n \leq (1 - 2\delta) K_1 \frac{d_{in} \log(p/d_{in})}{\log(\lambda)} \implies \inf_{\widehat{G}} \sup_{M \in \mathcal{M}_{p,d_{in}}(\lambda)} \Pr(\widehat{G} \neq G(M)) \geq \delta - \frac{K_2}{p d_{in} \log(p/d_{in})}.$$

□

Corollary B.2 (Gao et al., 2022). *Consider a subclass $\mathcal{G}' = \{G_1, ..., G_N\} \subseteq \mathcal{G}_{p,d_{in}}$, and let $\mathcal{M}' = \{M_1, ..., M_N\} \subseteq \mathcal{M}_{p,d_{in}}(\lambda)$, each of whose element is induced by one distinct $G \in \mathcal{G}'$. If a sample size is bounded as*

$$n \leq \frac{(1 - 2\delta) log N}{\alpha},$$

then any estimator for G is δ-unreliable:

$$\inf_{\widehat{G}} \sup_{M \in \mathcal{M}_{p,d_{in}}(\lambda)} \Pr(\widehat{G} \neq G(M)) \geq \delta - \frac{\log 2}{2 \log N},$$

where $\alpha \geq \max_{M_s \neq M_t \in \mathcal{M}'} KL(M_s \parallel M_t)$ and $G(M)$ denotes the DAG corresponding to M.

This lemma is exactly the same as Corollary B.2 of Gao et al. (2022), and its proof is directly from Lemma 3 of Yu (1997). Hence, the proof is omitted.

14 Numerical Experiments

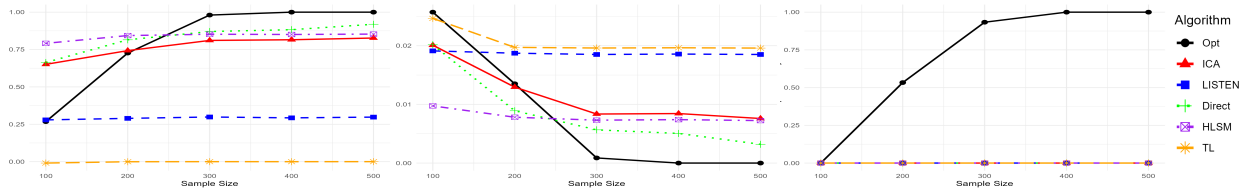
This section evaluates the numerical performance of Algorithm 1 (OptLiNGAM) and compares it with well-known LiNGAM learning methods, including ICALiNGAM (Shimizu et al., 2006), DirectLiNGAM (Shimizu et al., 2011), and TL (Zhao et al., 2022), as well as other linear SEM learning methods, such as LISTEN (Ghoshal and Honorio, 2018) and HLSM (Park et al., 2021). Specifically, OptLiNGAM utilizes the independence test based on distance covariance with significance level $\alpha = 2(1 - \Phi(\sqrt{c_0 n}))$, as shown in Theorem 5. The regularization parameters for LISTEN, HLSM, and TL, are implemented according to the recommended settings provided in the respective papers. For LISTEN algorithm, we set hard threshold $0.2 < \frac{\min\{|\beta_{j,k}| : \beta_{j,k} \neq 0\}}{2}$ and a sufficiently small regularized parameter $\lambda = 0.001$. For HLSM we set regularized parameter $\lambda \propto \sqrt{\frac{\log p}{n}}$. The graphical Lasso in TL is implemented using the default settings in the `R` package `TransGraph` (Zhao et al., 2022). Similarly, the significance level of independent tests in TL is set as $\alpha = 0.01$, as recommended in Zhao et al. (2022).

The simulation settings we consider are similar to those used by Wang and Drton (2020) and Zhao et al. (2022), where high-dimensional LiNGAM recovery was considered. In both settings, the data generation process is repeated 30 times.

- Setting 1

  Three types of hub graphs are generated with $d_{in} \in \{1, 2, 3\}$, where the number of hub nodes corresponds to d_{in}, $\lfloor \log p \rfloor$ nodes are isolated, and the remaining nodes are children of the hub nodes. Error terms are drawn from $\mathrm{Beta}(0.5, 0.5)$, with nonzero edge weights sampled uniformly from $[-1.5, -0.5] \cup [0.5, 1.5]$.

- Setting 2



(a) MCC (b) HM (c) Graph recovery rate

Figure 4: Average MCC, HM, and true graph recovery rate for 50-node sparse graphs with $d_{in} = 1$.

Graphs with $d_{in} = 1$ are generated, starting with a three-node graph featuring two directed edges from node 1 to nodes 2 and 3. At each step, a node is added with one directed edges, where the probability of establishing a directed edge from any existing node to the newly added node is contingent upon the number of neighbors of the existing node, thereby maintaining the fixed maximum indegree. Error terms are drawn from Beta$(0.5, 0.5)$, with nonzero edge weights sampled uniformly from $[-1.5, -1] \cup [1, 1.5]$.

All algorithms are evaluated using the Matthews correlation coefficient (MCC), the normalized structural Hamming distance (HM) (Tsamardinos et al., 2006), and the empirical probability of successful graph recovery. MCC measures the accuracy of the estimated directed edges, ranging from -1 to $+1$, where $+1$ indicates a perfect prediction, 0 represents an average random prediction, and -1 signifies an inverse prediction. Additionally, HM quantifies the similarity between the estimated and the true DAG, with lower HM values indicating better estimation accuracy.

14.1 Graph Structure Learning in Setting 2

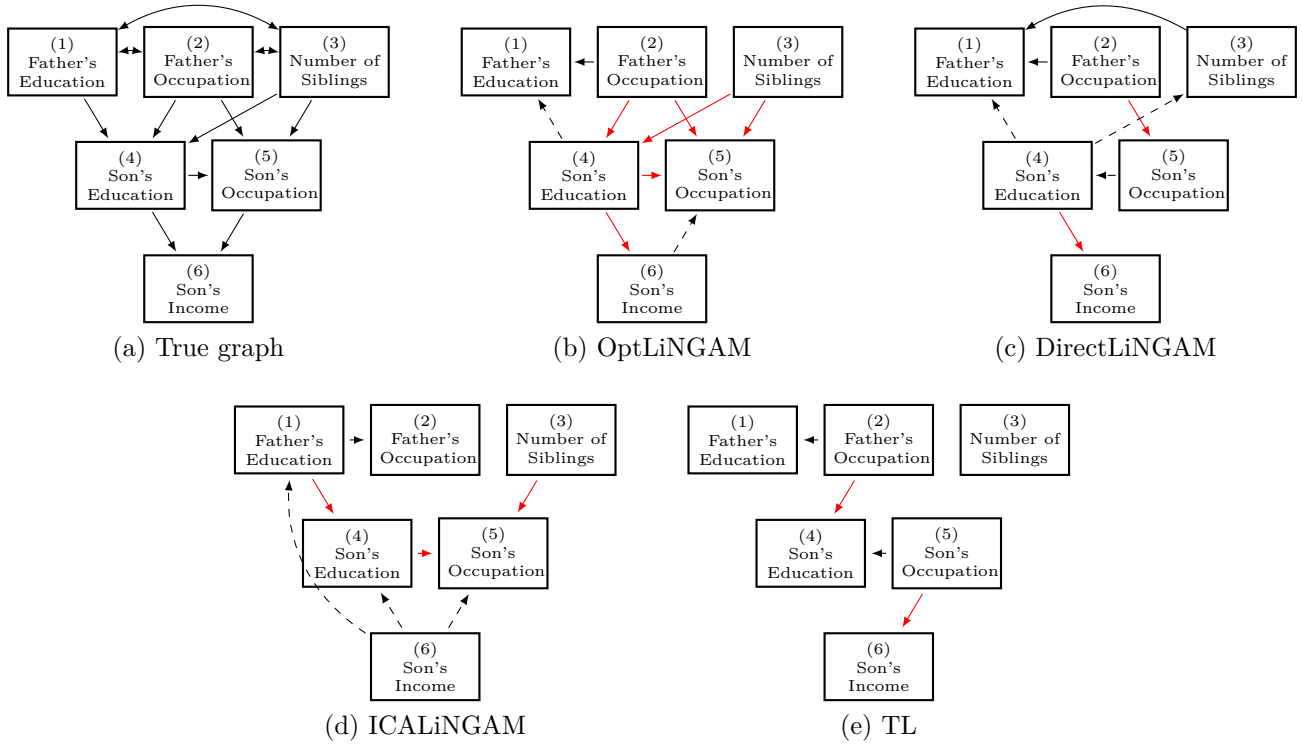
Figure 4 presents the average MCC, HM, and empirical probability of successful graph recovery for the algorithms considered in Setting 2, with sample sizes varying as $n \in \{100, 200, ..., 500\}$. As expected, all simulation results are analogous to those in Section 5.1, where a different type of model is recovered. Specifically, as n increases, the average MCC converges to 1, the average HM converges to 0, and the empirical probability of successful graph recovery converges to 1. Therefore, it validates the consistency of OptLiNGAM, as stated in Theorem 5.

Furthermore, Figure 4 compares the algorithms, demonstrating that OptLiNGAM significantly outperforms the comparison methods. Specifically, OptLiNGAM achieves higher MCC values, lower HM values, and higher graph recovery rates as the sample size increases. With more than 300 samples, OptLiNGAM exhibits superior performance in both MCC and HM compared to the other methods. Remarkably, only OptLiNGAM achieves consistent true graph recovery for all considered sample sizes. In contrast, no algorithm recovers the true graph even low-dimensional settings, due to the same reasons discussed in Section 5.1. These results confirm that OptLiNGAM reliably recovers the true graph in a limited-sample setting compared to other algorithms.

14.2 Graph Structure Learning in High-Dimensional Settings

Table 2: Empirical probability of successful graph recovery of all algorithms in high-dimensional settings. The best values are highlighted in bold.

(n, p)	**Method**	**Recovery Rate**	(n, p)	**Method**	**Recovery Rate**
(300,400)	Opt	**0.5333**	(300,500)	Opt	**0.3**
	ICA	0		ICA	0
	Direct	0		Direct	0
	TL	0		TL	0
	LISTEN	0		LISTEN	0
	HLSM	0		HLSM	0



(a) True graph (b) OptLiNGAM (c) DirectLiNGAM

(d) ICALiNGAM (e) TL

Figure 4: True and estimated status attainment graphs by the OptLiNGAM, DirectLiNGAM, ICALiNGAM, and TL algorithms.

To support the high-dimensional ($n < p$) consistency of the proposed algorithms as shown in Theorem 5, we evaluate the algorithms for recovering large-scale sparse LiNGAMs. Specifically, we focus on models from Setting 1 with $d_{in} = 1$ and nodes $p = 400$ and 500, using a fixed sample size of $n = 300$. Notably, the proposed method incurs a high-computational cost in exchange for optimal sample complexity as discussed in Section . Hence, learning large-scale models within a day may not always be achievable. If the graph recovery process exceeds 24 hours, the case is treated as a failure to recover the graph.

Table 2 shows the average graph recovery rates for the algorithms. Specifically, OptLiNGAM achieves empirical probabilities of successful graph recovery of 0.5333 and 0.3 for $p = 400$ and $p = 500$, respectively. In contrast, all other comparison methods fail to recover the true graph, owing to the same reason specified in Section 5.1. Overall, the simulation results heuristically confirm that OptLiNGAM consistently recovers the true graph even in high-dimensional settings. Its optimal sample efficiency is further highlighted when compared to existing methods in limited-sample settings.

15 Real Data Analysis

This section provides further real data analysis results. In the main manuscript, to confirm the strengths of our method, we apply the structure learning algorithms to a limited five-year data set, specifically from 2002 to 2006 with 355 observations. Here, we also consider seven other datasets, each derived by dividing the full data set from 1972 to 2006 into five-year intervals. Additionally, we compare OptLiNGAM with more algorithms, including not only DirectLiNGAM, but also ICALiNGAM and TL. To evaluate the strengths of each algorithm, we compare the Hamming distance, as well as the number of explainable and unexplainable edges based on domain knowledge.

Figure 4 presents the estimated graphs using the dataset from 2002 to 2006. In Figure 4, red directed edges indicate the explainable edges by domain knowledge and dashed directed edges represent relationships that are not consistent with domain knowledge. As discussed in Section 6, OptLiNGAM successfully identifies most causal relationships between variables, except for two reversed edges, (4, 1) and (6, 5). In contrast, DirectLiNGAM fails

to capture many important links, such as (2, 4) and (3, 5), which OptLiNGAM correctly identifies. Additionally it falsely determines the direction of edges, including (4, 3) and (5, 4), which are accurately detected by OptLiNGAM. ICALiNGAM correctly recovers the edge (1, 4), which is misdirected by OptLiNGAM. However, while ICALiNGAM detects the edges originating from (6), it misclassifies their direction and overlooks the edges that should arise from (2) and (3), which are all correctly identified by OptLiNGAM. Finally, while TL correctly recovers the edge (5, 6), which is also misdirected by OptLiNGAM. However, while TL detects the edge (5, 4), it incorrectly classifies its direction and overlooks many edges such as (2, 5), (3, 4), (3, 5) and (4, 6), all of which are accurately identified by OptLiNGAM.

Table 3 summarizes the performance of OptLiNGAM, DirectLiNGAM, ICALiNGAM, and TL in terms of the Hamming distance and the number of explainable and unexplainable edges using the dataset from 2002 to 2006. Specifically, OptLiNGAM outperforms the other algorithms with the lowest Hamming distance of 2, correctly identifying 6 explainable edges based on domain knowledge, while only introducing 2 unexplainable edges. In contrast, DirectLiNGAM, ICALiNGAM, and TL all show not only higher Hamming distances of 6, with DirectLiNGAM identifying only 2 explainable edges and introducing 3 unexplainable edges. ICALiNGAM recovers 3 explainable edges but also introduces 3 unexplainable edges, while TL identifies 2 explainable edges and introduces the fewest unexplainable edges among the three lower-performing methods. Overall, OptLiNGAM demonstrates superior performance in both accuracy and consistency with domain knowledge.

Table 3: Hamming distance and the number of (un)explainable edges in the estimated graphs using a reduced (2002-2006) data set.

	OptLiNGAM	**DirectLiNGAM**	**ICALiNGAM**	**TL**
Hamming distance	2	6	6	6
Explainable edges	6	2	3	2
Unexplainable edges	2	3	3	1

OptLiNGAM demonstrates strong performance not only with the specific dataset from 2002 to 2006 but also across multiple time intervals. Table 4 presents the average Hamming distance and the average number of explainable and unexplainable edges obtained by applying the algorithms to seven datasets, each derived by dividing the full dataset from 1972 to 2006 into five-year intervals.

As shown in Table 4, OptLiNGAM achieves the lowest average Hamming distance of 5.571, outperforming DirectLiNGAM (6.714), ICALiNGAM (7.000), and TL (7.286). Additionally, OptLiNGAM identifies the highest average number of explainable edges (3.857), compared to DirectLiNGAM (1.571), ICALiNGAM (1.429), and TL (0.857). However, OptLiNGAM also introduces the highest average number of unexplainable edges (4.000), followed by DirectLiNGAM (3.000), ICALiNGAM (2.429), and TL (1.289).

These results suggest a trade-off between type I and type II errors in independent testing, which can be adjusted by tuning the significance level α. The variability in the strength of dependence between variables across different time intervals may influence the value of ϵ in $\alpha = 2(1 - \Phi(\sqrt{n\epsilon}))$. Therefore, with appropriate adjustments, the performance of OptLiNGAM could potentially be improved further, balancing the number of explainable and unexplainable edges more effectively.

Table 4: Average Hamming distance and the number of (un)explainable edges across the estimated graphs from seven data sets of five-year intervals.

	OptLiNGAM	**DirectLiNGAM**	**ICALiNGAM**	**TL**
Hamming distance	5.571	6.714	7.000	7.286
Explainable edges	3.857	1.571	1.429	0.857
Unexplainable edges	4.000	3.000	2.429	1.289

References

Darmois, G. (1953). Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8.

Gao, M., Tai, W. M., and Aragam, B. (2022). Optimal estimation of gaussian dag models. In *International Conference on Artificial Intelligence and Statistics*, pages 8738–8757. PMLR.

Ghoshal, A. and Honorio, J. (2018). Learning linear structural equation models in polynomial time and sample complexity. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Huo, X. and Székely, G. J. (2016). Fast computing for distance covariance. *Technometrics*, 58(4):435–447.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Park, G., Moon, S. J., Park, S., and Jeon, J.-J. (2021). Learning a high-dimensional linear structural equation model via l1-regularized regression. *Journal of Machine Learning Research*, 22(102):1–41.

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

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

Skitovitch, V. P. (1953). On a property of the normal distribution. *DAN SSSR*, 89:217–219.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, Y. S. and Drton, M. (2020). High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1):41–59.

Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer.

Zhao, R., He, X., and Wang, J. (2022). Learning linear non-gaussian directed acyclic graph with diverging number of nodes. *Journal of Machine Learning Research*, 23(269):1–34.