Optimal Estimation of Gaussian (Poly)trees

Yuhao Wang
National University of Singapore

Ming Gao
University of Chicago

Wai Ming Tai
Nanyang Technological University

Bryon Aragam
University of Chicago

Arnab Bhattacharyya
National University of Singapore

Abstract

We develop optimal algorithms for learning undirected Gaussian trees and directed Gaussian polytrees from data. We consider both problems of distribution learning (i.e. in KL distance) and structure learning (i.e. exact recovery). The first approach is based on the Chow-Liu algorithm, and learns an optimal tree-structured distribution efficiently. The second approach is a modification of the PC algorithm for polytrees that uses partial correlation as a conditional independence tester for constraint-based structure learning. We derive explicit finite-sample guarantees for both approaches, and show that both approaches are optimal by deriving matching lower bounds. Additionally, we conduct numerical experiments to compare the performance of various algorithms, providing further insights and empirical evidence.

1 INTRODUCTION

Graphical models are a classical statistical tool for efficiently modeling data with rich, combinatorial structure. Directed acyclic graphs (DAGs) are widely used to capture causal relationships among complex systems. Probabilistic graphical models defined on DAGs, known as Bayesian networks (Pearl et al., 2000), have found broad applications in various disciplines, from biology (Markowetz and Spang, 2007; Zhang et al., 2013; Altay and Emmert-Streib, 2010), social science (Gupta and Kim, 2008), knowledge representation (Van Harmelen et al., 2008), data mining (Heckerman, 1997), recommendation systems (Hsu et al., 2012), legal decision making (Thagard, 2004), and more. When this structure is known in advance, it is straightforward to exploit this structure for inference tasks, among other things (Wainwright and Jordan, 2008). When this structure is unknown, it is must first be learned from data, which is the difficult problem of *structure learning* in graphical models. First, observational data only reveal the Markov equivalence class, captured by a completed partially directed acyclic graph (CPDAG Andersson et al., 1997). Classical approaches to learning a CPDAG from data include the PC algorithm (Spirtes and Glymour, 1991; Kalisch and Bühlman, 2007) and GES (Chickering, 2002; Nandy et al., 2018). Moreover, it is also known that the general problem of learning DAGs from observational data is an NP-complete problem (Chickering, 1996; Chickering et al., 2004; Chen et al., 2019), although a few polynomial-time algorithms have been proposed in special cases (Ghoshal and Honorio, 2017b; Chen et al., 2019; Park, 2020; Gao et al., 2020).

An important unresolved problem in this direction is to characterize the sample complexity of structure learning, or the minimum number of samples required to learn the graph from data. The past decade has produced a detailed theory for *undirected* graphical models (i.e. Markov random fields (Wainwright, 2019; Wang et al., 2010; Santhanam and Wainwright, 2012)). By comparison, much less is known about DAGs. In this paper, we study in detail the simplest unresolved DAG model, namely, directed Gaussian trees. Perhaps surprisingly, despite its simplicity, and unlike in the undirected case, the optimal sample complexity of learning directed Gaussian trees has remained an open problem. Suppose we are given sample access to a Gaussian distribution $P = \mathcal{N}(0, \Sigma)$, where the goal is to learn a DAG G that represents P. While we defer formal definitions to Section 2, we can broadly summarize three different problems:

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1. (Non-realizable setting) When P is an arbitrary Gaussian (i.e. not representable by any tree), how many samples are required to learn a tree-structured distribution Q that is optimally close to P?
2. (Realizable setting) When P itself is tree-structured, how samples are required to learn a tree-structured distribution Q that is optimally close to P?
3. (Faithful setting) When P is faithful to some tree T, how samples are required to learn T itself (i.e. the tree structure) up to Markov equivalence?

It is well-known that each of these problems is solvable—in principle—under different assumptions. For example, the celebrated Chow-Liu algorithm solves the first two problems, however, whether or not this can be improved with a more efficient algorithm is unknown. The same goes for the third setting: The famous PC and GES algorithms can find a faithful DAG (even without the tree assumption), however, their optimality remains unresolved. One of our main contributions is to study all three problems in a single unified setting, allowing for apples-to-apples comparisons of the assumptions required, and the resulting (optimal) sample complexity for each. See Section 6 for more along these lines.

Although faithfulness can be a strong assumption in practice, we emphasize that to the best of our knowledge, no optimality results under this assumption are known. Thus, our analysis presents a possible first foray in this direction. Previous work has shown that faithfulness is notoriously challenging to analyze (e.g. Uhler et al., 2013; Gao et al., 2023).

1.1 Our Contributions

We are given n i.i.d. samples $X = (X^{(1)}, \dots, X^{(n)}) \in \mathbb{R}^{n\times d}$ from an unknown Gaussian P. We consider two distinct but canonical problems: *Distribution learning* and *structure learning*. The difference between these two problems lies in the error metric: In distribution learning, we seek to learn P in KL-divergence, with no respect for underlying structure (i.e. there may be no structure at all), whereas in structure learning, we assume *a priori* the existence of a tree T and seek to learn T exactly, with no respect for the distribution P. Structure learning is known to require restrictive assumptions, and thus part of our effort is to illustrate how different assumptions lead to different conclusions and sample complexities. With this in mind, our results consider three progressively stronger assumptions on P: Non-realizable, realizable, and faithful.

Below, we outline our main contributions at a high-level, while deferring precise statements and problem formulations to Section 3 and Section 4.

Non-realizable Setting Without making additional assumptions on P, we show that*

$$n = \widetilde{\Theta}\Big(\frac{d^2}{\varepsilon^2}\Big) \tag{1.1}$$

samples are necessary and sufficient to learn (with probability at least 2/3) a tree-structured distribution that is ε-close to the closest tree-structured distribution for P.

Realizable Setting When P itself is Markov to a tree T (i.e. it is *tree-structured*), then

$$n = \widetilde{\Theta}\Big(\frac{d}{\varepsilon}\Big) \tag{1.2}$$

samples are necessary and sufficient to learn (with probability at least 2/3) a tree-structured distribution that is ε-close to P itself.

Faithful Polytrees Switching our goal from learning the closest tree-structured distribution to structure learning, we additionally assume that P is faithful to some *polytree* T. We show that the optimal sample complexity of learning $\overline{T}$, the CPDAG of T, is

$$n = \Theta\Big(\frac{\log d}{c^2}\Big), \tag{1.3}$$

where c is a faithfulness parameter defined in (4.2).

Clearly, and unsurprisingly, realizable distribution learning is easier than the non-realizable case. A more interesting question is how to compare these to structure learning. In Section 6, we conclude with a discussion and comparison of these two cases, with some intriguing directions for future work.

1.2 Other Related Work

Learning Bayesian Networks Bayesian network structure learning has been extensively studied, and the reader may consult one of several overviews for more details and background (Spirtes et al., 2000; Pearl et al., 2000; Koller and Friedman, 2009; Murphy, 2012; Peters et al., 2017; Maathuis et al., 2018; Squires and Uhler, 2022). Classical approaches assume faithfulness, a condition that permits learning of the Markov equivalence class, such as constraint-based methods (Spirtes and Glymour, 1991; Friedman et al., 2013) and score-based approaches (Chickering, 2002; Nandy et al., 2018). A different strand of research has explored a range of alternative distributional assumptions that allow for effective learning such as non-gaussianity (Shimizu et al.,

*$\widetilde{\Theta}$ is used to ignore potential log factors.

2006; Shimizu, 2014; Wang and Drton, 2020), nonlinearity (Hoyer et al., 2008; Zhang and Hyvärinen, 2009) or equal error variances (Peters and Bühlmann, 2014; Ghoshal and Honorio, 2017b, 2018; Chen et al., 2019; Gao et al., 2020).

When it comes to the tree-structured models, the classical Chow-Liu algorithm (Chow and Liu, 1968; Chow and Wagner, 1973) can recover the skeleton of a non-degenerate polytree in the equivalence class. Furthermore, One of the first papers to consider the problem of learning polytrees was Rebane and Pearl (1987), after which Dasgupta (1999) showed that learning polytrees is NP-hard in general. Srebro (2003) has shown that the related problem of finding the maximum likelihood graphical model with bounded treewidth is also NP-hard. Recently Tan et al. (2010, 2011) investigated the difficulty of learning trees and forests, while Liu et al. (2011) adopted a nonparametric approach using kernel density estimates. The Chow-Liu algorithm has also been applied for learning latent locally tree-like graphs (Anandkumar and Valluvan, 2013).

Sample Complexity of Structure Learning Early work to consider the sample complexity problem for Bayesian networks includes Friedman and Yakhini (1996); Zuk et al. (2012). More recently, for distribution learning over finite alphabets, Daskalakis and Pan (2020, 2021) showed that d-variable tree-structured Ising models can be learned computationally-efficiently to within total variation distance ε from an optimal $O(d \log d/\varepsilon^2)$ samples. Around the same time, Bhattacharyya et al. (2023) derived explicit sample complexity bounds for the Chow-Liu algorithm of $\widetilde{O}(d\varepsilon^{-1})$ for trees on d vertices, and $d^2\varepsilon^{-2}$ samples for a general distribution P. Choo et al. (2023) further extend Bhattacharyya et al. (2023) into d-polytree when the underlying graph skeleton is known.

The literature on structure learning is comparatively deeper; however, it has traditionally forgone concerns about optimality and lower bounds. As this is our main focus, we focus here on prior work on optimal algorithms. Ghoshal and Honorio (2017a) first established lower bounds for a range of DAG models, after which Gao et al. (2022) showed that a variant of the algorithm from Chen et al. (2019) achieves optimal sample complexity of $n \asymp q \log(d/q)$ for equal variance DAGs (Peters and Bühlmann, 2014; Loh and Bühlmann, 2014), where q is the maximum number of parents and d is the number of nodes. To the best of our knowledge, optimality results and lower bounds in the faithful setting are missing, one exception is the sub-problem of neighbourhood selection (Gao et al., 2023), and one of our main contributions is to partially fill this gap. We mention prior work that considers consistency and upper bounds under faithfulness (Kalisch and Bühlman, 2007; Nandy et al., 2018; Rothenhäusler et al., 2018), relaxation and improvement on classical methods (Chickering, 2020; Marx et al., 2021; Lam et al., 2022), and recent progress on learning polytrees (Gao and Aragam, 2021; Azadkia et al., 2021; Tramontano et al., 2022; Jakobsen et al., 2022).

Furthermore, developing a (conditional) independence tester with respect to mutual information with $o(1/\varepsilon^2)$ sample complexity was posed as an open problem in Canonne et al. (2018). Canonne et al. (2018) showed that both Ising model *goodness-of-fit testing* and *independence testing* can be solved from $\mathrm{poly}(d, 1/\varepsilon)$ samples in polynomial time. More details related to the *distribution property testing* can be found in Rubinfeld (2012); Canonne (2020); Goldreich (2017); Bhattacharyya and Yoshida (2022).

2 PRELIMINARIES AND TOOLS

Preliminary Notions We employ standard asymptotic notation $O(\cdot), \Omega(\cdot), \Theta(\cdot)$; and as usual, $\widetilde{\cdot}$ indicates up to log factors. For example, if $f = \widetilde{\Theta}(g)$ then $f = O(g(\log g)^{c_1})$ and $f = \Omega(g/(\log g)^{c_2})$ for some constants c_1 and c_2. We say $f \lesssim g$ and $f \gtrsim g$ if $f \leq Cg$ and $f \geq cg$ for some positive constants C and c.

Graphical Definitions For a directed acyclic graph (DAG) $G = (V, E)$, for each node $k \in V$, $\mathrm{pa}(k) = \{j : (j,k) \in E\}$ denotes its parent nodes, descendants $\mathrm{de}(k)$ denotes the nodes that can be reached by k and $\mathrm{nd}(k) = V \setminus \mathrm{de}(k)$ denotes the nondescendants. The skeleton of G, $\mathrm{sk}(G)$, is the undirected graph formed by removing directions of all the edges in G. For any $j, \ell, k \in V$, a triple (j, ℓ, k) is called unshielded if both j, k are adjacent to ℓ but not adjacent to each other, graphically $j - \ell - k$; and is called a v-structure if additionally j, k are parents of ℓ, i.e. $j \to \ell \leftarrow k$. The in-degree of G is $\max_k |\mathrm{pa}(k)|$. A tree is an undirected graph in which any two nodes are connected by exactly one path. A directed tree is a directed graph in which, for some root node u, and any other node v, there is exactly one directed path from u to v. A polytree is a directed graph whose skeleton to be a tree. Denote the set of directed trees (resp. polytrees) over d nodes to be $\mathcal{T}$ (resp. $\widetilde{\mathcal{T}}$). Note that a directed tree is a polytree with in-degree equal to one except the root node who has no parent and $\mathcal{T} \subseteq \widetilde{\mathcal{T}}$.

Gaussian Bayesian Networks Given a random vector $X = (X_1, \ldots, X_d)$ drawn from a distribution P, a DAG G is a Bayesian network for X (or precisely, its joint distribution P) if the following factorization

holds:

$$P(X) = \prod_{k=1}^{d} P(X_k \,|\, X_{\mathrm{pa}(k)})\,. \tag{2.1}$$

Here, we use $X = V = [d]$ interchangeably with some abuse of notation. From now on, we assume that $P = \mathcal{N}(0, \Sigma)$ throughout. Since P is Gaussian, we can always express X as the following linear structural equation model (SEM):

$$X_k = \beta_k^\top X + \eta_k\,, \quad \eta_k \sim \mathcal{N}(0, \sigma_k^2), \tag{2.2}$$

where $\beta_k \in \mathbb{R}^d$ is supported on $\mathrm{pa}(k)$ and the $\{\eta_k\}_{k=1}^d$ are mutually independent. A Gaussian distribution is said to be T-structured for some directed tree $T \in \mathcal{T}$ (or simply tree-structured when the specific T is not important in the context) if it satisfies (2.1) with respect to some tree T. For a distribution P and a directed tree T, let

$$P_T := \operatorname*{arg\,min}_{T\text{-structured distribution } Q} D_{\mathrm{KL}}(P \,\|\, Q),$$

where $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ denotes the KL-divergence. In this paper, we consider both general Gaussians (non-realizable case) as well as tree-structured distributions (realizable and faithful cases), i.e. (2.2) holds for some directed (poly)tree T.

Faithfulness and Markov Equivalence Class For the purpose of structure learning, a common assumption is faithfulness, under which the DAG is identified up to its Markov equivalence class (MEC). We assume the reader is familiar with standard graphical concepts such as d-separation; see (Koller and Friedman, 2009) for more background.

Definition 2.1 (Faithfulness)**.** *We say a distribution P is faithful to a DAG G if for any $j, k \in V$ and $S \subseteq V \setminus \{j, k\}$,*

$$X_j \perp\!\!\!\perp X_k \,|\, X_S \Rightarrow j \text{ and } k \text{ are } d\text{-separated by } S\,.$$

Equivalently, for any two nodes j and k not d-separated by set S, faithfulness requires $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$. The MEC of a DAG G is the set of DAGs that encode the same set of conditional independencies as G, which is usually represented by a CPDAG, denoted by $\overline{G}$. A standard approach to learning a CPDAG under faithfulness is the PC algorithm (Spirtes and Glymour, 1991), which relies on conditional independence testing to recover the skeleton and orient the edges. While faithfulness can be a strong assumption (Uhler et al., 2013), it is known that weaker assumptions suffice. For example:

Definition 2.2 (Restricted faithfulness)**.** *We say a distribution P is restricted faithful to a DAG G if*

1. *For any $(j, k) \in E$, $S \subseteq V \setminus \{j, k\}$, $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$;*
2. *For any unshielded triple $j - \ell - k$, if this is a v-structure, then $X_j \not\perp\!\!\!\perp X_k \,|\, S$ for any $S \subseteq V \setminus \{j, k\}$ with $\ell \in S$; if not, then $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$ for any $S \subseteq V \setminus \{j, k, \ell\}$.*

Under general faithfulness, all conditional independence relationships imply d-separations in a DAG. In other words, all instances of d-connections lead to conditional dependence. On the contrary, restricted faithfulness requires only a subset of d-connections to imply conditional dependence. Conventionally, the first part of Definition 2.2 is also named *adjacency-faithfulness* and the second part is named *orientation-faithfulness*. With our focus on the setup where the underlying DAG is a polytree, restricted faithfulness can be further relaxed as we will discuss in Section 4.

3 LEARNING TREE-STRUCTURED GAUSSIANS

We begin by studying the sample complexity for learning tree-structured Gaussian distributions. For any $\varepsilon > 0$, we would like to devise an algorithm taking samples drawn from a Gaussian P that returns a directed tree $\widehat{T} \in \mathcal{T}$ and a distribution $P_{\widehat{T}}$ that is Markov to $\widehat{T}$ such that

$$D_{\mathrm{KL}}(P \,\|\, P_{\widehat{T}}) \le \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \,\|\, P_T) + \varepsilon\,,$$

We seek to achieve this goal with a minimal number of samples. Notably, for any $T \in \mathcal{T}$, $D_{\mathrm{KL}}(P \,\|\, P_T)$ can be expressed as

$$-\sum_{i=1}^{d} I(X_i; X_{\mathrm{pa}(i)}) - H(X) + \sum_{i=1}^{d} H(X_i), \tag{3.1}$$

where H is the entropy function and I is the mutual information.

3.1 Distribution Learning Upper Bounds

The classical Chow-Liu algorithm (Chow and Liu, 1968) builds the maximum weight spanning tree where the weight of the "potential" edge between nodes j and k is the estimated mutual information $\widehat{I}(X_j, X_k)$ from data. Although its return is an undirected graph, we modify the output to be any directed tree whose skeleton matches the undirected graph with light abuse of notation. This is because any $T \in \mathcal{T}$ with the same skeleton will share the same P_T, which is the target of distribution learning analyzed in the sequel.

Our first result gives an upper bound on the sample complexity for distribution learning in the non-realizable setting:

Algorithm 1: Modified `Chow-Liu` algorithm

Input: n i.i.d. samples $(X_1^{(i)}, \dots, X_d^{(i)})$

1. For each $j = 1, \dots, d$:
   (a) $\widehat{\sigma}_j^2 \leftarrow \frac{1}{n}\sum_{i=1}^n (X_j^{(i)})^2$
2. For each pair $(j,k), 1 \le j < k \le d$:
   (a) $\widehat{\rho}_{jk} \leftarrow \frac{1}{n}\sum_{i=1}^n X_j^{(i)} X_k^{(i)}$
3. For each pair $(j,k), 1 \le j < k \le d$:
   (a) $\widehat{I}(X_j; X_k) \leftarrow -\frac{1}{2}\log\left(1 - \frac{\widehat{\rho}_{jk}^2}{\widehat{\sigma}_j^2\widehat{\sigma}_k^2}\right)$ which is same as $\frac{1}{2}\log(1 + \frac{\widehat{\beta}_{jk}^2\widehat{\sigma}_j^2}{\widehat{\sigma}_{k|j}})$ defined in Section B.2
4. $G \leftarrow$ the weighted complete undirected graph on $[d]$ whose edge weight for (j,k) is $\widehat{I}(X_j; X_k)$
5. $\widehat{S} \leftarrow$ the maximum weighted spanning tree of G
6. $\widehat{T} \leftarrow$ any directed tree with skeleton to be $\widehat{S}$

Output: A directed tree $\widehat{T}$

Theorem 3.1. *Let P be a Gaussian distribution. Given n i.i.d. samples from P, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d^2}{\varepsilon^2}\log\frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 1 satisfies*

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \le \min_{T\in\mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon,$$

with probability at least $1-\delta$.

When P is Markov to a tree (i.e. it is tree-structured), then the sample complexity improves:

Theorem 3.2. *Let T^* be a directed tree and P_{T^*} be a T^*-structured Gaussian. Given n i.i.d. samples from P_{T^*}, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d}{\varepsilon}\log\frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 1 satisfies*

$$D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}}) \le \varepsilon,$$

with probability at least $1-\delta$.

Remark: We can also obtain a sample-efficient algorithm for bounded-degree Gaussian *polytrees*, using the guarantees of the estimator $\widehat{I}$, assuming that the skeleton is known. We defer the description of this result to Appendix B.5.

3.2 Distribution Learning Lower Bounds

The main idea of our proof is to reduce a distribution testing problem to our problem. Intuitively, the distribution testing problem is defined as follows. Suppose $R^{(1)}$ and $R^{(2)}$ are two distributions whose $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)})$ is small. We are given n i.i.d. samples drawn from a distribution P where P is a m-variate distribution and each coordinate is distributed as either $R^{(1)}$ or $R^{(2)}$ uniformly and independently. Our task is to determine which of $R^{(1)}$ or $R^{(2)}$ the samples are drawn from correctly for at least $m/2$ coordinates. The formal definition will be presented in Problem B.7. When $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)})$ is sufficiently small, one should expect that n needs to be large enough to solve this problem with probability 2/3. Hence, we construct the $(R^{(1)}, R^{(2)})$ pairs for the non-realizable and realizable case accordingly.

Theorem 3.3. *Suppose P is an unknown Gaussian distribution. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d^2/\varepsilon^2)$, no algorithm returns a directed tree $\widehat{T}$ such that*

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \le \min_{T\in\mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon$$

with probability at least 2/3.

Theorem 3.4. *Suppose P is an unknown Gaussian distribution such that there exists a directed tree T^* that P is T^*-structured, i.e. $P = P_{T^*}$. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d/\varepsilon)$, no algorithm returns a directed tree $\widehat{T}$ such that*

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \le \varepsilon$$

with probability at least 2/3.

4 OPTIMAL FAITHFUL TREE LEARNING

In the preceding section, we learned a tree-structured distribution under the KL distance, without concern for the learned tree structure. This viewpoint primarily pertains to *distribution learning*. This section adopts an different approach, emphasizing the aspect of *structure learning*. Specifically, we assume the underlying graph structure is indeed a tree, more generally, a polytree. We introduce an estimator based on the classic PC algorithm (Spirtes and Glymour, 1991) and analyze its sample complexity under faithfulness. Crucially, we provide a matching lower bound to conclude the minimax optimality of the algorithm, which offers insights into the difficulty of structure learning under faithfulness.

4.1 Tree-Faithfulness

As alluded to in Section 2, the tree structure allows us to relax the usual notion of faithfulness:

Definition 4.1 (Tree-faithfulness)**.** *We say distribution P is tree-faithful to a polytree T if*

1. *For any two nodes connected $X_j - X_k$, we have $X_k \not\perp\!\!\!\perp X_j \,|\, X_\ell$ for all $\ell \in V \cup \{\emptyset\} \setminus \{k, j\}$;*
2. *For any v-structure $X_k \to X_\ell \leftarrow X_j$, we have $X_k \not\perp\!\!\!\perp X_j \,|\, X_\ell$.*

Tree-faithfulness comprises two components, each corresponding to adjacency-faithfulness and orientation-faithfulness respectively in restricted faithfulness (cf. Definition 2.2). In comparison to adjacency-faithfulness, tree-faithfulness solely requires conditional dependence for neighbouring nodes with conditioning sets of size at most one. Likewise, compared to orientation faithfulness, tree-faithfulness only needs conditional dependence for v-structures given the the collider. Let $\rho(X_j, X_k \,|\, X_\ell)$ be the conditional correlation coefficient between X_k and X_j given X_ℓ. As usual, in order to establish uniform, finite-sample results, we need the following concept of c-strong tree-faithfulness:

Definition 4.2 (c-strong tree-faithfulness)**.** *We say that P is c-strong tree-faithful to a polytree T if*

1. *For any two nodes connected $X_j - X_k$, we have $\rho(X_k, X_j \,|\, X_\ell) \geq c$ for $\ell \in V \cup \{\emptyset\} \setminus \{k, j\}$;*
2. *For any v-structure $X_k \to X_\ell \leftarrow X_j$, we have $\rho(X_k, X_j \,|\, X_\ell) \geq c$.*

Under strong tree-faithfulness, we can now establish how the sample complexity depends on both the dimension d and the signal strength c.

4.2 Structure Learning Upper Bounds

We develop the `PC-Tree` algorithm for learning polytrees as a modification to the classic PC algorithm, outlined in Algorithm 2, effectively identifying the polytree's skeleton. An important by-product is the separation set resulted from the CI testing, which is used to obtain the CPDAG by applying an `ORIENT` step (Algorithm 3) as in the original PC algorithm.

In contrast to the original PC algorithm, `PC-Tree` distinguishes itself in two key aspects. Firstly, when assessing the presence of an edge between any two nodes, instead of exploring all potential conditioning sets, `PC-Tree` simplifies the process by exclusively testing marginal independence and conditional independence given only one other node. Furthermore, a notable departure from the original PC algorithm is that `PC-Tree` combines marginal independence tests and conditional independence tests, as opposed to ignoring the latter once marginal independence is established. `PC-Tree` will rely on sample (conditional) correlation coefficient for all the (conditional) independence tests when running the algorithm, see more details in Appendix C.1.

Algorithm 2: `PC-Tree` algorithm

Input: n i.i.d. samples $(X_1^{(i)}, \ldots, X_d^{(i)})$

1. Let $\widehat{E} = \emptyset$.
2. For each pair (j, k), $0 \leq j < k \leq d$:
   (a) For all $\ell \in [d] \cup \{\emptyset\} \setminus \{j, k\}$:
      i. Test $H_0 : X_j \perp\!\!\!\perp X_k \,|\, X_\ell$ vs. $H_1 : X_j \not\perp\!\!\!\perp X_k \,|\, X_\ell$, store the results.
   (b) If all tests reject, then $\widehat{E} \leftarrow \widehat{E} \cup \{j - k\}$.
   (c) Else (if some test accepts), let $S(j, k) = \{\ell \in [d] \cup \{\emptyset\} \setminus \{j, k\} : X_j \perp\!\!\!\perp X_k \,|\, X_\ell\}$.

Output: $\widehat{T} = ([d], \widehat{E})$, separation set S.

Now we are ready to provide the sample complexity of `PC-Tree` in the following theorem, whose proof is postponed to Appendix C.2 and C.3.

Theorem 4.3. *For any $T \in \widetilde{\mathcal{T}}$, assuming P is c-strong tree-faithful to T, applying Algorithm 2 with sample correlation for CI testing, if the sample size*

$$n \gtrsim \frac{1}{c^2}\Big(\log d + \log(1/\delta)\Big),$$

then $\Pr(\widehat{T} = \mathrm{sk}(T)) \geq 1 - \delta$, *and* $\Pr(\textsc{Orient}(\widehat{T}, S) = \overline{T}) \geq 1 - \delta$

We may compare this upper bound $(\log d)/c^2$ with some of existing results on structure learning. Compared to learning equal variance general DAGs (Gao et al., 2022) with optimal rates being $q \log(d/q)$, tree structure simplifies the problem by removing the factor of in-degree q. As against recovering undirected graph in MRF (Misra et al., 2020), whose optimal sample complexity is $(s \log d)/\kappa^2$, we are able to improve the rate by the maximum degree s. Moreover, considering directed trees $T \in \mathcal{T} \subset \widetilde{\mathcal{T}}$, Lemma A.1 shows c to be a constant under mild assumption on the parametrization of (2.2), which assures possible concern of dependence on c.

4.3 Structure Learning Lower Bounds

Having provided the sample complexity upper bound, we continue to derive a matching lower bound:

Theorem 4.4. *Assuming c-strong tree-faithfulness, and $c^2 \leq 1/5$, $d \geq 4$, if the sample size is bounded as*

$$n \leq \frac{1 - 2\delta}{8} \times \frac{\log d}{c^2},$$

then for any estimator $\widehat{T}$ for $\overline{T}$,

$$\inf_{\widehat{T}} \sup_{\substack{T \in \widetilde{\mathcal{T}} \\ P \text{ is } c\text{-strong} \\ \text{tree-faithful to } T}} \Pr(\widehat{T} \neq \overline{T}) \geq \delta - \frac{\log 2}{\log d}.$$

The lower bound in Theorem 4.4 implies the optimal sample complexity is $\Theta(\log d/c^2)$, where the dependence on $1/c^2$ term characterizes the hardness from "how (Tree-)faithful" the distribution is; and $\log d$ term comes from the cardinality of all polytrees, which is much smaller compared to number of all DAGs.

To prove this lower bound, we employee Fano's inequality (Yu, 1997) and consider a subclass of $\mathcal{T}$ to exploit the property that any node in directed tree has at most one parent. This subclass of directed trees has large enough cardinality by Cayley's formula of undirected trees. With the parametrization of edge weights appropriately calibrated, we show the KL divergence between the distributions consistent with any two instances from the subclass is well controlled, which leads to the final lower bound. The detailed proof can be found in Appendix C.4.

Remark 4.5. *The optimality results in this section also extend to directed tree, polyforest and Markov chain. Since the lower bound is constructed using directed trees, the optimality applies. For polyforest, which is essentially polytree but allows for disconnected component,* `PC-Tree` *algorithm is able to identify the correct skeleton. On the other hand, polytree is a subclass of polyforest, thus the lower bound in Theorem 4.4 applies. For Markov chain, the algorithm is modified to dismiss marginal independence test, and the lower bound construction considers all Markov chains with the same way of parametrization as in Theorem 4.4. All these graphical models share the optimal sample complexity* $\Theta(\log d/c^2)$.

5 EXPERIMENTS

We conduct experiments to verify our findings in structure learning. For brevity, we report here only the most difficult setting with $d = 100$ nodes; full details on the experiments and additional setups, e.g. when noise η_k is not Gaussian, can be found in Appendix D. We simulated random directed trees and synthetic data via (2.2). We compare the performance of PC-Tree, Chow-Liu to PC and GES as classical baselines when only faithfulness assumed. Though Chow-Liu algorithm aims for distribution learning, it also estimates the skeleton as a byproduct. Therefore, to make fair comparison, we evaluate them by the accuracy of skeleton of the outputs (of PC-Tree, PC and GES). The results on average Structural Hamming Distance (SHD) and the Precise Recovery Rate (PRR) are reported in Figure 1, where PRR measures the relative frequency of exact recovery. From the figure, we can see PC-Tree algorithm does perform the best, especially the significantly better result on PRR over the baselines, which is the main metric we are concerned with and have established optimality for. The competitive performance of Chow-Liu is also noticeable, for which we have not analyzed under the goal of structure learning, and we conjecture a similar sample complexity is shared with PC-Tree.

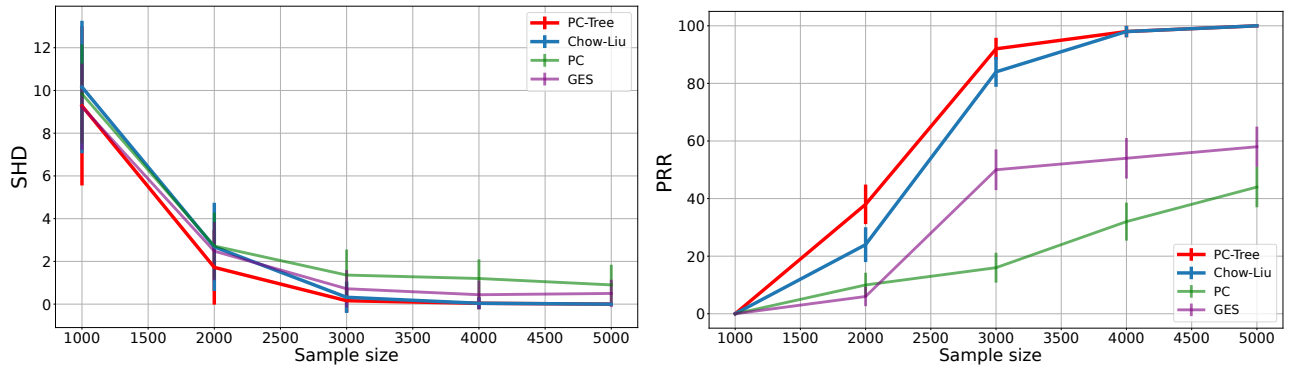
6 COMPARISON AND DISCUSSION

The literature on distribution learning and structure learning have largely evolved separate from one another. An interesting aspect of our results is that they treat both problems in a unified setting, allowing for an explicit comparison of these problems.

First, it is clear that the non-realizable setting should not be compared to structure learning, since in the former setting there is no structure to speak of. In the realizable setting, however, it is reasonable to ask for a comparison. Comparing (1.2) and (1.3), it is easy to see that there is a phase transition when $\varepsilon \asymp dc^2$. Focusing on directed trees for an apple-to-apple comparison, if the SEM parameters, e.g. β_k, σ_k^2 in (2.2) are bounded, then strong tree-faithfulness holds with $c \asymp 1$, see Lemma A.1. In this case, the optimal sample complexity is $\log d$ for structure learning and $(d \log d)/\varepsilon$ for distribution learning, which has an additional factor of d/ε. Thus, as long as $\varepsilon = o(d)$, which is typical, structure learning is easier than distribution learning.

Another interesting scenario arises when $\varepsilon \ll dc^2$: Here, distribution learning is harder, however, we might hope to learn the structure of T "for free" by first learning the distribution to within KL accuracy ε. This is because, as ε goes to zero, $\widehat{P}$ converges to P, which implies we can use $\widehat{P}$ directly to estimate partial correlations for structure learning. Then the question boils down to whether there exists a good estimator of the structure that exploits $\widehat{P}$ when $\varepsilon \ll dc^2$. Lemma A.2 shows that as long as the estimator is agnostic to $\widehat{P}$ (in the sense that it treats $\widehat{P}$ as a black-box input), then we must have at least $\varepsilon \ll c^2$. Thus, there is a regime $c^2 \ll \varepsilon \ll dc^2$ where distribution learning does not automatically imply structure learning, at least in general. It remains as an interesting open question how small ε must be for $\widehat{P}$ to be efficiently used for structure learning, or whether or not there exist *specific* estimators $\widehat{P}$ that can be used for structure learning when $c^2 \ll \varepsilon \ll dc^2$.

Extending these results beyond the Gaussians we consider here (as well as finite alphabets as in previous work) is a promising direction for future research. Especially interesting would be bounds in a non-parametric setting.



(a) SHD comparison

(b) PRR comparison

Figure 1: Performance comparison for PC-Tree, Chow-Liu, PC and GES algorithm evaluated on SHD and PRR. The red, blue, green, purple lines are for PC-Tree, Chow-Liu, PC and GES respectively.

Acknowledgements

BA was supported by NSF IIS-1956330, NIH R01GM140467, and the Robert H. Topel Faculty Research Fund at the University of Chicago Booth School of Business. WT was supported by the Singapore MOE AcRF Tier 2 grant MOE-T2EP20122-0001. This work was done in part while YW, BA, and AB were visiting the Simons Institute for the Theory of Computing.

References

Altay, G. and Emmert-Streib, F. (2010). Inferring the conservative causal core of gene regulatory networks. *BMC systems biology*, 4(1):1–13.

Anandkumar, A. and Valluvan, R. (2013). Learning loopy graphical models with latent variables: Efficient methods and guarantees. *The Annals of Statistics*, pages 401–435.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*, volume 2. Wiley New York.

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541.

Azadkia, M., Taeb, A., and Bühlmann, P. (2021). A fast non-parametric approach for local causal structure learning. *arXiv preprint arXiv:2111.14969*.

Bhattacharyya, A., Gayen, S., Price, E., Tan, V. Y., and Vinodchandran, N. (2023). Near-optimal learning of tree-structured distributions by chow and liu. *SIAM Journal on Computing*, 52(3):761–793.

Bhattacharyya, A. and Yoshida, Y. (2022). *Property Testing: Problems and Techniques*. Springer Nature.

Canonne, C. L. (2020). A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100.

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. (2018). Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 735–748. ACM.

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

Chickering, D. M. (1996). Learning bayesian networks is np-complete. *Learning from data: Artificial intelligence and statistics V*, pages 121–130.

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

Chickering, M. (2020). Statistically efficient greedy equivalence search. In *Conference on Uncertainty in Artificial Intelligence*, pages 241–249. PMLR.

Chickering, M., Heckerman, D., and Meek, C. (2004). Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330.

Choo, D., Yang, J. Q., Bhattacharyya, A., and Canonne, C. L. (2023). Learning bounded-degree polytrees with known skeleton. *arXiv preprint arXiv:2310.06333*.

Chow, C. and Wagner, T. (1973). Consistency of an estimate of tree-dependent probability distributions (corresp.). *IEEE Transactions on Information Theory*, 19(3):369–371.

Chow, C. K. and Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory*, 14(3):462–467.

Dasgupta, S. (1999). Learning polytrees. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 134–141.

Daskalakis, C. and Pan, Q. (2020). Tree-structured ising models can be learned efficiently. *arXiv preprint arXiv:2010.14864*.

Daskalakis, C. and Pan, Q. (2021). Sample-optimal and efficient learning of tree ising models. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 133–146.

Friedman, N., Nachman, I., and Pe'er, D. (2013). Learning bayesian network structure from massive datasets: The" sparse candidate" algorithm. *arXiv preprint arXiv:1301.6696*.

Friedman, N. and Yakhini, Z. (1996). On the sample complexity of learning bayesian networks. In *Uncertainty in Artifical Intelligence (UAI)*.

Gao, M. and Aragam, B. (2021). Efficient bayesian network structure learning via local markov boundary search. *Advances in Neural Information Processing Systems*, 34:4301–4313.

Gao, M., Ding, Y., and Aragam, B. (2020). A polynomial-time algorithm for learning nonparametric causal graphs. *Advances in Neural Information Processing Systems*, 33:11599–11611.

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.

Gao, M., Tai, W. M., and Aragam, B. (2023). Optimal neighbourhood selection in structural equation models. *arXiv preprint arXiv:2306.02244*.

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

Ghoshal, A. and Honorio, J. (2017b). Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. *Advances in Neural Information Processing Systems*, 30.

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

Goldreich, O. (2017). *Introduction to property testing*. Cambridge University Press.

Gupta, S. and Kim, H. W. (2008). Linking structural equation modeling to bayesian networks: Decision support for customer retention in virtual communities. *European Journal of Operational Research*, 190(3):818–833.

Heckerman, D. (1997). Bayesian networks for data mining. *Data mining and knowledge discovery*, 1:79–119.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.

Hsu, F.-M., Lin, Y.-T., and Ho, T.-K. (2012). Design and implementation of an intelligent recommendation system for tourist attractions: The integration of ebm model, bayesian network and google maps. *Expert Systems with Applications*, 39(3):3257–3264.

Jakobsen, M. E., Shah, R. D., Bühlmann, P., and Peters, J. (2022). Structure learning for directed trees. *The Journal of Machine Learning Research*, 23(1):7076–7172.

Kalisch, M. and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3).

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lam, W.-Y., Andrews, B., and Ramsey, J. (2022). Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR.

Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J., and Wasserman, L. (2011). Forest density estimation. *The Journal of Machine Learning Research*, 12:907–951.

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.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks–a review. *BMC bioinformatics*, 8(6):1–17.

Marx, A., Gretton, A., and Mooij, J. M. (2021). A weaker faithfulness assumption based on triple interactions. In *Uncertainty in Artificial Intelligence*, pages 451–460. PMLR.

Misra, S., Vuffray, M., and Lokhov, A. Y. (2020). Information theoretic optimal learning of gaussian graphical models. In *Conference on Learning Theory*, pages 2888–2909. PMLR.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

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

Park, G. (2020). Identifiability of additive noise models using conditional variances. *The Journal of Machine Learning Research*, 21(1):2896–2929.

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.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Rebane, G. and Pearl, J. (1987). The recovery of causal poly-trees from statistical data. *Uncertainty in Artificial Intelligence'87*, pages 222–228.

Rothenhäusler, D., Ernest, J., Bühlmann, P., et al. (2018). Causal inference in partially linear structural equation models. *The Annals of Statistics*, 46(6A):2904–2938.

Rubinfeld, R. (2012). Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24–28.

Santhanam, N. P. and Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134.

Shimizu, S. (2014). Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41:65–98.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72.

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

Squires, C. and Uhler, C. (2022). Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, pages 1–35.

Srebro, N. (2003). Maximum likelihood bounded tree-width markov networks. *Artificial intelligence*, 143(1):123–138.

Tan, V. Y., Anandkumar, A., and Willsky, A. S. (2010). Learning gaussian tree models: Analysis of error exponents and extremal structures. *IEEE Transactions on Signal Processing*, 58(5):2701–2714.

Tan, V. Y., Anandkumar, A., and Willsky, A. S. (2011). Learning high-dimensional markov forest distributions: Analysis of error rates. *Journal of Machine Learning Research*, 12:1617–1653.

Thagard, P. (2004). Causal inference in legal decision making: Explanatory coherence vs. bayesian networks. *Applied Artificial Intelligence*, 18(3-4):231–249.

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

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*, 41(2):436–463.

Van Harmelen, F., Lifschitz, V., and Porter, B. (2008). *Handbook of knowledge representation*. Elsevier.

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

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Wang, W., Wainwright, M. J., and Ramchandran, K. (2010). Information-theoretic bounds on model selection for gaussian markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE.

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, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., Zhang, C., Xie, T., Tran, L., Dobrin, R., et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707–720.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press.

Zuk, O., Margel, S., and Domany, E. (2012). On the number of samples needed to learn the correct structure of a bayesian network.

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]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (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]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

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]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (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]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [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. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A COMPARING STRUCTURE LEARNING AND DISTRIBUTION LEARNING

Lemma A.1. *Suppose $T \in \mathcal{T}$ and P is parameterized using $\{\beta_k, \sigma_k^2\}_{k=1}^d$ as (2.2) according to T. If there exists a constant $M > 1$ such that for any $k \in [d]$,*

$$|\beta_{kj}| \in [M^{-1}, M], \qquad \forall \beta_{kj} \neq 0$$
$$\sigma_k^2 \in [M^{-1}, M],$$

then P is c-strong Tree-faithful to T for some $c \asymp 1$.

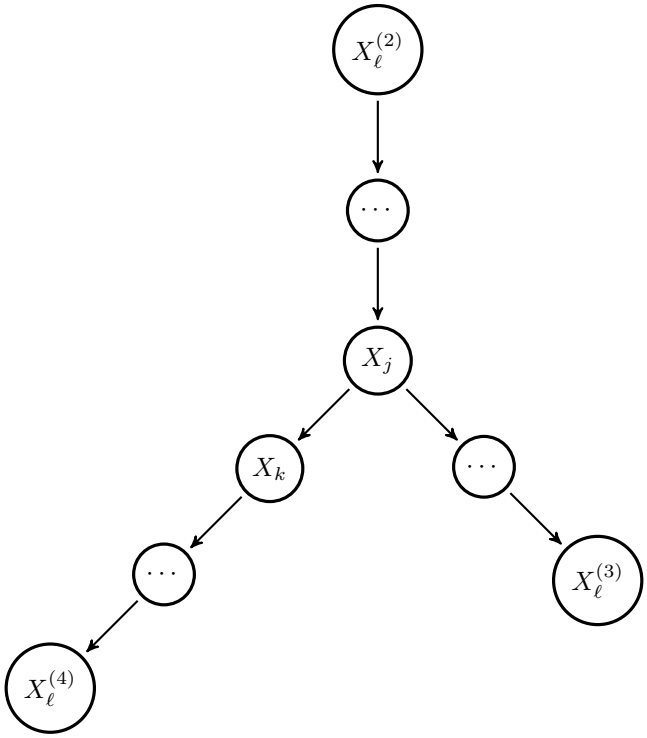


Figure 2: Four cases of ℓ to verify for c-strong Tree-faithfulness, indicated by the superscript of X_ℓ. The first case is when $\ell = \emptyset$. The second, third and fourth are when ℓ is the ancestor of j, descendant of j and descendant of k.

Proof of Lemma A.1. Since a directed tree T does not have any v-structures, we only need to verify adjacency faithfulness in Definition 4.2. For any two nodes connected as $j \to k$, we want to check whether $\rho(X_j, X_k \,|\, X_\ell)$ is lower bounded by some constant for $\ell \in V \cup \{\emptyset\} \setminus \{j, k\}$. There are four cases of ℓ to consider, see Figure 2:

- $\ell = \emptyset$: To simplify the notation, we write

$$X_k = \beta_k \times X_j + \eta_k$$

with $\beta_k \in \mathbb{R}$ and $|\beta_k| \in [M^{-1}, M]$, $\mathrm{var}(\eta_k) = \sigma_k^2$. We also write $V_j^2 := \mathrm{var}(X_j) \geq \sigma_j^2$. Hence,

$$\rho(X_j, X_k) = \frac{\beta_k V_j^2}{\sqrt{V_j^2}\sqrt{\beta_k^2 V_j^2 + \sigma_k^2}} = \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2 V_j^2}} \geq \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2\sigma_j^2}} \gtrsim 1\,.$$

- $\ell \in \mathrm{an}(j)$: Write $V_{j\,|\,\ell}^2 = \mathrm{var}(X_j \,|\, X_\ell) \geq \sigma_j^2$, hence

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{\beta_k V_{j\,|\,\ell}^2}{\sqrt{V_{j\,|\,\ell}^2}\sqrt{\beta_k^2 V_{j\,|\,\ell}^2 + \sigma_k^2}} \geq \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2\sigma_j^2}} \gtrsim 1\,.$$

- $\ell \in \mathrm{de}(j)$: Suppose the directed path from j to ℓ is $j \to h_1 \to h_2 \to \ldots \to h_q \to \ell$, q can be 0, then we can write

$$X_\ell = b_1 X_j + u_1\,,$$

with

$$b_1 = \beta_\ell \prod_{i=1}^{q} \beta_{h_i}, \qquad u_1 = \eta_\ell + \beta_\ell \sum_{i=1}^{q} \eta_{h_i} \prod_{t=i+1}^{q} \beta_{h_t}\,,$$

and

$$\nu_1^2 := \mathrm{var}(u_1) = \sigma_\ell^2 + \beta_\ell^2 \sum_{i=1}^{q} \sigma_{h_i}^2 \prod_{t=i+1}^{q} \beta_{h_t}^2 \geq \beta_\ell^2 \sigma_{h_1}^2 \prod_{t=2}^{q} \beta_{h_t}^2\,.$$

So we have $b_1^2/\nu_1^2 \leq \beta_{h_1}^2/\sigma_{h_1}^2 \asymp 1$. The covariance among X_j, X_k, X_ℓ is

$$\mathrm{cov}(X_j, X_k, X_\ell) = \begin{pmatrix} V_j^2 & \beta_k V_j^2 & b_1 V_j^2 \\ * & \beta_k^2 V_j^2 + \sigma_k^2 & b_1\beta_k V_j^2 \\ * & * & b_1^2 V_j^2 + \nu_1^2 \end{pmatrix}.$$

Then the conditional covariance is

$$\mathrm{cov}(X_j, X_k \,|\, X_\ell) \propto \begin{pmatrix} \nu_1^2 & \beta_k \nu_1^2 \\ * & \beta_k^2\nu_1^2 + \sigma_k^2 b_1^2 + \sigma_k^2\nu_1^2/V_j^2 \end{pmatrix}.$$

Therefore,

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{1}{\sqrt{1 + \frac{\sigma_k^2}{\beta_k^2} \times \frac{b_1^2}{\nu_1^2} + \frac{\sigma_k^2}{V_j^2\beta_k^2}}} \gtrsim 1\,.$$

- $\ell \in \mathrm{de}(k)$: Similarly, we can write

$$X_\ell = b_2 X_k + u_2, \qquad \mathrm{var}(u_2) = \nu_2^2\,,$$

with $b_2^2/\nu_2^2 \lesssim 1$. The covariance among X_j, X_k, X_ℓ is

$$\mathrm{cov}(X_j, X_k, X_\ell) = \begin{pmatrix} V_j^2 & \beta_k V_j^2 & b_2\beta_k V_j^2 \\ * & V_k^2 & b_2 V_k^2 \\ * & * & b_2^2 V_k^2 + \nu_2^2 \end{pmatrix}.$$

Then the conditional covariance is

$$\mathrm{cov}(X_j, X_k \,|\, X_\ell) \propto \begin{pmatrix} b_2^2\sigma_k^2 V_j^2 + \nu_2^2 V_j^2 & \beta_k V_j^2 \nu_2^2 \\ * & \nu_2^2 V_k^2 \end{pmatrix}.$$

Therefore,

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{1}{\sqrt{(1 + \frac{\sigma_k^2}{\beta_k^2 V_j^2})(1 + \frac{b_2^2}{\nu_2^2}\sigma_k^2)}} \gtrsim 1\,.$$

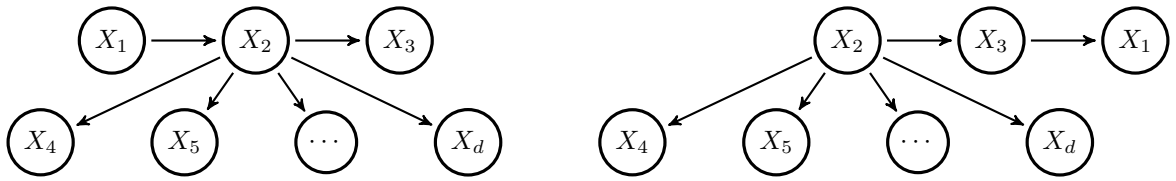
In all four cases, $\rho(X_j, X_k \mid X_\ell) \gtrsim 1$, thus c-strong Tree-faithfulness is satisfied with some $c \asymp 1$. □

Lemma A.2. *Let $\mathcal{A}$ denote some distribution learning algorithm such that given a tree-structured distribution P, $\mathcal{A}$ takes data from P and outputs $\widehat{P}$ with $D_{\mathrm{KL}}(P\|\widehat{P}) \le \varepsilon$. If $\varepsilon \gtrsim c^2$, then for any estimator $\widehat{T}(\widehat{P})$ for $\overline{T}$ using solely $\widehat{P}$,*

$$\inf_{\widehat{T}(\widehat{P})} \sup_{\substack{T\in\mathcal{T} \\ P \text{ is } c\text{-strong} \\ \text{Tree-faithful to } T}} \sup_{\mathcal{A}} \Pr(\widehat{T}(\widehat{P}) \neq \overline{T}) = 1\,.$$

Proof. We construct $T, T' \in \mathcal{T}$ with different skeletons, and P, P' Markov and strongly faithful to T, T' respectively such that $D_{\mathrm{KL}}(P\|P') \asymp c^2$. In this way, consider the ground truth to be T and P, and supppose $\mathcal{A}$ outputs $\widehat{P} = P'$. Then we have $D_{\mathrm{KL}}(P\|\widehat{P}) \le \varepsilon$ with $\varepsilon \asymp c^2$. While P and $\widehat{P} = P'$ correspond to different structures, thus any estimator using solely $\widehat{P}$ cannot uniformly find the true structure.

It remains to show the construction: Consider T and T' as follows:



(a) Tree T (b) Tree T'

Figure 3: Construction for Lemma A.2.

We parameterize P, P' as the lower bound construction in Appendix C.4:

$$X_k = \beta X_{\mathrm{pa}(k)} + \eta_k\,,$$

where $\beta = \sqrt{2}c$, $\eta_k \sim \mathcal{N}(0,1)$ and Lemma C.5 makes sure they are c-strong tree faithful. Now we only need to compute the KL divergence:

$$\begin{aligned}
D_{\mathrm{KL}}(P\|P') &= \mathop{\mathbb{E}}_{P} \log \frac{\prod_k P(X_k \mid \mathrm{pa}(k))}{\prod_j P'(X_j \mid \mathrm{pa}(j))} \\
&= \mathop{\mathbb{E}}_{P} \log \frac{P(X_3 \mid X_2)P(P_2 \mid X_1)P(X_1)}{P(X_1 \mid X_3)P(X_3 \mid X_2)P(X_2)} \\
&= \mathop{\mathbb{E}}_{P} \frac{1}{2}\Big(X_2^2 + (X_1 - \beta X_3)^2 - X_1^2 - (X_3 - \beta X_2)^2 \Big) \\
&= \frac{1}{2}\Big(-\beta^4 + \beta^6 + 2(\beta^2 + \beta^4 - \beta^3) \Big) \\
&\le 2\beta^2 = 4c^2\,,
\end{aligned}$$

which completes the proof. □

B PROOFS OF Section 3

B.1 Preliminaries

We first state some useful lemmas. They are well-known results for the concentration bound on variances and covariances. For completeness, we provide the proof below.

Lemma B.1 (Guarantees of variance recovery)**.** *Suppose X is the random variable of $\mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. Let $X^{(1)}, \ldots, X^{(n)}$ be the i.i.d. samples of X and $\widehat{\sigma}^2$ be $\frac{1}{n}\sum_{i=1}^n (X^{(i)})^2$. Then, for any $t \in (0,1)$, we have*

$$|\widehat{\sigma}^2 - \sigma^2| < t\sigma^2$$

with probability $1 - O(e^{-\Omega(nt^2)})$.

Proof. We first show that the probability of $\widehat{\sigma}^2 > (1+t)\sigma^2$ is bounded by $e^{-\Omega(nt^2)}$ and the other inequality $\widehat{\sigma}^2 < (1-t)\sigma^2$ follows similarly.

Note that

$$\widehat{\sigma}^2 > (1+t)\sigma^2 \Leftrightarrow e^{\lambda\frac{1}{n}\sum_{i=1}^n (X^{(i)})^2} > e^{\lambda(1+t)\sigma^2} \qquad \text{for any } \lambda > 0.$$

By Markov inequality, the probability of $\widehat{\sigma}^2 > (1+t)\sigma^2$ is bounded by

$$\mathbb{E}(e^{\lambda\frac{1}{n}\sum_{i=1}^n (X^{(i)})^2})/e^{\lambda(1+t)\sigma^2} = \underbrace{\mathbb{E}(e^{\lambda\frac{1}{n}X^2})^n}_{\text{by i.i.d. assumption}} /e^{\lambda(1+t)\sigma^2}. \tag{B.1}$$

Hence, we need to bound the term $\mathbb{E}(e^{\lambda\frac{1}{n}X^2})$.

$$\mathbb{E}(e^{\lambda\frac{1}{n}X^2}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\lambda\frac{1}{n}x^2} e^{-\frac{1}{2\sigma^2}x^2} dx = \frac{1}{\sqrt{1-\frac{2\sigma^2\lambda}{n}}} \qquad \text{as long as } \frac{1}{2\sigma^2} - \frac{\lambda}{n} > 0$$

Moreover, using the inequality $\frac{1}{\sqrt{1-x}} \le e^{\frac{1}{2}x+x^2}$ for $x < \frac{1}{2}$, we have

$$\mathbb{E}(e^{\lambda\frac{1}{n}X^2}) \le e^{\frac{\sigma^2\lambda}{n}+\frac{4\sigma^4\lambda^2}{n^2}} \qquad \text{as long as } \frac{2\sigma^2\lambda}{n} < \frac{1}{2} \tag{B.2}$$

Plugging (B.2) into (B.1), the probability of $\widehat{\sigma}^2 > (1+t)\sigma^2$ is bounded by

$$(e^{\frac{\sigma^2\lambda}{n}+\frac{4\sigma^4\lambda^2}{n^2}})^n/e^{\lambda(1+t)\sigma^2} = e^{-\frac{4\sigma^4\lambda^2}{n}+\lambda t\sigma^2} = e^{-\frac{4\sigma^4}{n}(\lambda-\frac{nt}{8\sigma^2})^2+\frac{nt^2}{16}}$$

and, by taking $\lambda = \frac{nt}{8\sigma^2}$, it becomes $e^{-\frac{nt^2}{16}}$.

□

Lemma B.2 (Guarantees of correlation coefficient recovery)**.** *Suppose (X,Y) is the random variable of $\mathcal{N}(0,\Sigma)$ for some positive definite $\Sigma = \begin{bmatrix}\sigma_x^2 & \rho_{xy}\\ \rho_{xy} & \sigma_y^2\end{bmatrix}$. Let $(X^{(1)},Y^{(1)}),\dots,(X^{(n)},Y^{(n)})$ be the i.i.d. samples of (X,Y) and $\widehat{\rho}_{xy}$ be $\frac{1}{n}\sum_{i=1}^n X^{(i)}Y^{(i)}$. Then, for any $t \in (0,1)$, we have*

$$|\widehat{\rho}_{xy} - \rho_{xy}| < t\sigma_x\sigma_y$$

with probability $1 - O(e^{-\Omega(nt^2)})$.

Proof. We first show that the probability of $\widehat{\rho}_{xy} > \rho_{xy} + t\sigma_x\sigma_y$ is bounded by $e^{-\Omega(nt^2)}$ and the other inequality $\widehat{\rho}_{xy} < \rho - t\sigma_x\sigma_y$ follows similarly.

Note that

$$\widehat{\rho}_{xy} > \rho_{xy} + t\sigma_x\sigma_y \Leftrightarrow e^{\lambda\frac{1}{n}\sum_{i=1}^n X^{(i)}Y^{(i)}} > e^{\lambda(\rho_{xy}+t\sigma_x\sigma_y)} \qquad \text{for any } \lambda > 0.$$

By Markov inequality, the probability of $\widehat{\rho}_{xy} > \rho_{xy} + t\sigma_x\sigma_y$ is bounded by

$$\mathbb{E}(e^{\lambda\frac{1}{n}\sum_{i=1}^n X^{(i)}Y^{(i)}})/e^{\lambda(\rho_{xy}+t\sigma_x\sigma_y)} = \underbrace{\mathbb{E}(e^{\lambda\frac{1}{n}XY})^n}_{\text{by i.i.d. assumption}} /e^{\lambda(\rho_{xy}+t\sigma_x\sigma_y)}. \tag{B.3}$$

Hence, we need to bound the term $\mathbb{E}(e^{\lambda\frac{1}{n}XY})$.

$$\begin{aligned}\mathbb{E}(e^{\lambda\frac{1}{n}XY}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^2(\sigma_x^2\sigma_y^2-\rho_{xy}^2)}} e^{\lambda\frac{1}{n}xy} e^{-\frac{1}{2(\sigma_x^2\sigma_y^2-\rho_{xy}^2)}(\sigma_y^2x^2-2\rho_{xy}xy+\sigma_x^2y^2)} dxdy\\ &= \frac{1}{\sqrt{1-\frac{2\rho_{xy}\lambda}{n}-\frac{\lambda^2\Delta}{n^2}}} \qquad \text{as long as } \sigma_x^2\sigma_y^2 > (\rho_{xy}+\frac{\lambda\Delta}{n})^2 \text{ where } \Delta = \sigma_x^2\sigma_y^2-\rho_{xy}^2\end{aligned}$$

Moreover, using the inequality $\frac{1}{\sqrt{1-x}} \le e^{\frac{1}{2}x+x^2}$ for $x < \frac{1}{2}$, we have

$$\begin{aligned}
\mathbb{E}(e^{\lambda\frac{1}{n}XY}) &\le e^{\frac{1}{2}(\frac{2\rho_{xy}\lambda}{n}+\frac{\lambda^2\Delta}{n^2})+(\frac{2\rho_{xy}\lambda}{n}+\frac{\lambda^2\Delta}{n^2})^2} \qquad \text{as long as } \frac{2\rho_{xy}\lambda}{n}+\frac{\lambda^2\Delta}{n^2} < \frac{1}{2} \\
&\le e^{\frac{\rho_{xy}\lambda}{n}+\frac{\lambda^2\sigma_x^2\sigma_y^2}{2n^2}+(\frac{2\sigma_x\sigma_y\lambda}{n}+\frac{\lambda^2\sigma_x^2\sigma_y^2}{n^2})^2} \qquad \text{using } \rho_{xy} \le \sigma_x\sigma_y \text{ and } \Delta \le \sigma_x^2\sigma_y^2 \\
&\le e^{\frac{\rho_{xy}\lambda}{n}+\frac{19\lambda^2\sigma_x^2\sigma_y^2}{2n^2}} \qquad \text{as long as } \frac{\lambda\sigma_x\sigma_y}{n} < 1
\end{aligned} \tag{B.4}$$

Plugging (B.4) into (B.3), the probability of $\widehat{\rho}_{xy} > \rho_{xy} + t\sigma_x\sigma_y$ is bounded by

$$(e^{\frac{\rho_{xy}\lambda}{n}+\frac{19\lambda^2\sigma_x^2\sigma_y^2}{2n^2}})^n / e^{\lambda(\rho_{xy}+t\sigma_x\sigma_y)} = e^{-\frac{19\sigma_x^2\sigma_y^2}{2n}\lambda^2+t\sigma_x\sigma_y\lambda} = e^{-\frac{19\sigma_x^2\sigma_y^2}{2n}(\lambda-\frac{tn}{19\sigma_x\sigma_y})^2+\frac{t^2n}{38}}$$

and, by taking $\lambda = \frac{tn}{19\sigma_x\sigma_y}$, it becomes $e^{-\frac{t^2n}{38}}$.

□

Corollary B.3. *Suppose $(X_1, \dots, X_d)$ is the random variable of $\mathcal{N}(0, \Sigma)$ for some positive definite Σ where $\rho_{ij} := \Sigma_{ij}$ and $\sigma_i^2 := \Sigma_{ii}$ for $i, j = 1, \dots, d$. Let $(X_1^{(1)}, \dots, X_d^{(1)}), \dots, (X_1^{(n)}, \dots, X_d^{(n)})$ be the i.i.d. samples of $(X_1, \dots, X_d)$ and*

$$\widehat{\rho}_{jk} = \frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} X_k^{(i)} \qquad \text{and} \qquad \widehat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} (X_j^{(i)})^2.$$

Then, when $n = \Theta(\frac{1}{t^2}\log\frac{d}{\delta})$, we have, for all $j, k = 1, \dots, d$,

$$|\widehat{\rho}_{jk} - \rho_{jk}| \le t\sigma_j\sigma_k \qquad \text{and} \qquad |\widehat{\sigma}_j^2 - \sigma_j^2| \le t\sigma_j^2$$

with probability $1 - \delta$.

B.2 Conditional Mutual Information Tester

In this subsection, we define the conditional mutual information tester used in our main algorithm.

Suppose (X, Y, Z) is the random variable of $\mathcal{N}(0, \Sigma)$ for some positive definite $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho_{xy} & \rho_{xz} \\ \rho_{xy} & \sigma_y^2 & \rho_{yz} \\ \rho_{xz} & \rho_{xy} & \sigma_z^2 \end{bmatrix}$. WLOG, we can express (X, Y, Z) as

$$\begin{aligned}
Y &= \beta_{xy}X + \eta_y \\
Z &= \gamma_{xz}X + \gamma_{yz}Y + \eta_z
\end{aligned}$$

for some random variables η_y, η_z where

$$\beta_{xy} = \frac{\rho_{xy}}{\sigma_x^2} \qquad \text{and} \qquad \begin{bmatrix} \gamma_{xz} \\ \gamma_{yz} \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho_{xy} \\ \rho_{xy} & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} \rho_{xz} \\ \rho_{yz} \end{bmatrix}.$$

Let $\sigma_{y|x}^2$ be $\mathbb{E}(\eta_y^2)$ and $\sigma_{z|x,y}^2$ be $\mathbb{E}(\eta_z^2)$. Recall that the mutual information $I(X;Y)$ and the conditional mutual information $I(Y;Z \mid X)$ are defined (equivalently) as

$$I(X;Y) := \frac{1}{2}\log(1 + \frac{\beta_{xy}^2\sigma_x^2}{\sigma_{y|x}^2}) \qquad \text{and} \qquad I(Y;Z \mid X) := \frac{1}{2}\log(1 + \frac{\gamma_{yz}^2\sigma_{y|x}^2}{\sigma_{z|x,y}^2})$$

Let $(X^{(1)}, Y^{(1)}, Z^{(1)}), \dots, (X^{(n)}, Y^{(n)}, Z^{(n)})$ be the i.i.d. samples of (X, Y, Z). Then we define the empirical mutual information $\widehat{I}(X;Y)$ and the empirical mutual information $\widehat{I}(Y;Z \mid X)$ to be

$$\widehat{I}(X;Y) := \frac{1}{2}\log(1 + \frac{\widehat{\beta}_{xy}^2\widehat{\sigma}_x^2}{\widehat{\sigma}_{y|x}^2}) \qquad \text{and} \qquad \widehat{I}(Y;Z \mid X) := \frac{1}{2}\log(1 + \frac{\widehat{\gamma}_{yz}^2\widehat{\sigma}_{y|x}^2}{\widehat{\sigma}_{z|x,y}^2}) \tag{B.5}$$

where the $\widehat{\cdot}$ mark indicates the empirical version of the quantity. Namely,

$$\begin{cases} \widehat{\sigma}_x^2 & := \frac{1}{n}\sum_{i=1}^n (X^{(i)})^2, \quad \widehat{\sigma}_y^2 := \frac{1}{n}\sum_{i=1}^n (Y^{(i)})^2, \quad \widehat{\sigma}_z^2 := \frac{1}{n}\sum_{i=1}^n (Z^{(i)})^2, \\ \widehat{\rho}_{xy} & := \frac{1}{n}\sum_{i=1}^n X^{(i)}Y^{(i)}, \quad \widehat{\rho}_{xz} := \frac{1}{n}\sum_{i=1}^n X^{(i)}Z^{(i)}, \quad \widehat{\rho}_{yz} := \frac{1}{n}\sum_{i=1}^n Y^{(i)}Z^{(i)}, \\ \widehat{\beta}_{xy} & := \frac{\widehat{\rho}_{xy}}{\widehat{\sigma}_x^2}, \quad \begin{bmatrix} \widehat{\gamma}_{xz} \\ \widehat{\gamma}_{yz} \end{bmatrix} := \begin{bmatrix} \widehat{\sigma}_x^2 & \widehat{\rho}_{xy} \\ \widehat{\rho}_{xy} & \widehat{\sigma}_y^2 \end{bmatrix}^{-1} \begin{bmatrix} \widehat{\rho}_{xz} \\ \widehat{\rho}_{yz} \end{bmatrix}, \\ \widehat{\sigma}_{y|x}^2 & := \widehat{\sigma}_y^2 - \widehat{\beta}_{xy}^2 \widehat{\sigma}_x^2 \quad \text{and} \quad \widehat{\sigma}_{z|x,y}^2 := \widehat{\sigma}_z^2 - \widehat{\gamma}_{xz}^2 \widehat{\sigma}_x^2 - \widehat{\gamma}_{yz}^2 \widehat{\sigma}_{y|x}^2. \end{cases} \tag{B.6}$$

Note that the above quantities depend on the samples but we will not emphasize it if the set of samples is clear in the context. Also, it is known that, by the chain rule of mutual information,

$$I(X;Y) - I(X;Z) = I(X;Y \mid Z) - I(X;Z \mid Y) \tag{B.7}$$

$$\widehat{I}(X;Y) - \widehat{I}(X;Z) = \widehat{I}(X;Y \mid Z) - \widehat{I}(X;Z \mid Y). \tag{B.8}$$

From now on, when we have a d-dimensional random variable $(X_1, \ldots, X_d)$, we abuse the notations defined in (B.6) by replacing x, y, z with i, j, k for $i, j, k = 1, \ldots, d$.

Lemma B.4. *Suppose $(X_1, \ldots, X_d)$ is the random variable of $\mathcal{N}(0, \Sigma)$ for some positive definite Σ where $\rho_{ij} := \Sigma_{ij}$ and $\sigma_i^2 := \Sigma_{ii}$ for $i, j = 1, \ldots, d$. Let $(X_1^{(1)}, \ldots, X_d^{(1)}), \ldots, (X_1^{(n)}, \ldots, X_d^{(n)})$ be the i.i.d. samples of $(X_1, \ldots, X_d)$ and $\widehat{\gamma}_{ij}, \widehat{\sigma}_{i|j}, \widehat{\sigma}_{i|j,k}$ be the quantities defined in (B.6) for $i, j, k = 1, \ldots, d$. Then, when $n = \Theta(\frac{1}{t^2}\log\frac{d}{\delta})$, we have, for all $i, j, k = 1, \ldots, d$,*

$$|\widehat{\gamma}_{ij} - \gamma_{ij}| < t\frac{\sigma_{j|i,k}}{\sigma_{i|k}}, \qquad |\widehat{\sigma}_{i|j}^2 - \sigma_{i|j}^2| < t\sigma_{i|j}^2 \qquad \text{and} \qquad |\widehat{\sigma}_{i|j,k}^2 - \sigma_{i|j,k}^2| < t\sigma_{i|j,k}^2$$

with probability $1 - \delta$.

Proof. By using Corollary B.3 and the definition in (B.6), it can be done by a straightforward calculation. □

Theorem B.5 (Conditional Mutual Information Tester)**.** *Suppose $(X_1, \ldots, X_d)$ is the random variable of $\mathcal{N}(0, \Sigma)$ for some positive definite Σ. Let $(X_1^{(1)}, \ldots, X_d^{(1)}), \ldots, (X_1^{(n)}, \ldots, X_d^{(n)})$ be the i.i.d. samples of $(X_1, \ldots, X_d)$ For any sufficiently small $\varepsilon, \delta > 0$, if*

$$n = \Theta(\frac{1}{\varepsilon}\log\frac{d}{\delta}),$$

the following results hold for all $i, j, k = 1, \ldots, d$ with probability $1 - \delta$:

1. *If $I(X_i; X_j \mid X_k) = 0$, then $\widehat{I}(X_i; X_j \mid X_k) \le \frac{\varepsilon}{100}$.*
2. *If $I(X_i; X_j \mid X_k) \ge \varepsilon$, then $\widehat{I}(X_i; X_j \mid X_k) > \frac{1}{20}I(X_i; X_j \mid X_k) - \frac{\varepsilon}{40}$.*

Combining these two cases, we have

$$\widehat{I}(X_i; X_j \mid X_k) > \frac{1}{20}I(X_i; X_j \mid X_k) - \frac{\varepsilon}{40}.$$

Proof. By Lemma B.4, with $\Theta(\frac{1}{\varepsilon}\log\frac{d}{\delta})$, we have the following properties for all $i, j, k = 1, \ldots, d$ with probability $1 - \delta$:

$$|\widehat{\gamma}_{ij} - \gamma_{ij}| < \frac{\sqrt{\varepsilon}}{100}\frac{\sigma_{j|i,k}}{\sigma_{i|k}}, \qquad |\widehat{\sigma}_{i|j}^2 - \sigma_{i|j}^2| < \frac{\sqrt{\varepsilon}}{100}\sigma_{i|j}^2 \qquad \text{and} \qquad |\widehat{\sigma}_{i|j,k}^2 - \sigma_{i|j,k}^2| < \frac{\sqrt{\varepsilon}}{100}\sigma_{i|j,k}^2 \tag{B.9}$$

We express

$$\widehat{I}(X_i; X_j \mid X_k) = \frac{1}{2}\log\left(1 + \widehat{\gamma}_{ij}^2\frac{\widehat{\sigma}_{i|k}^2}{\widehat{\sigma}_{j|i,k}^2}\right) = \frac{1}{2}\log\left(1 + \widehat{\gamma}_{ij}^2\frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2}\cdot\frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2}\cdot\frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2}\right) \tag{B.10}$$

We bound each term $\widehat{\gamma}_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2}$, $\frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2}$ and $\frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2}$ for the cases of $I(X_i; X_j \mid X_k) = 0$ and $I(X_i; X_j \mid X_k) \geq \varepsilon$.

We first prove if $I(X_i; X_j \mid X_k) = 0$ then $\widehat{I}(X_i; X_j \mid X_k) \leq \frac{\varepsilon}{100}$. Since $I(X_i; X_j \mid X_k) = 0$, it means that X_i and X_j are independent conditioned on X_k and hence $\gamma_{ij} = 0$. We have $\widehat{\gamma}_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \leq \frac{\varepsilon}{100}$. For the term $\frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2}$, we have $\frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2} \leq 1 + \frac{\sqrt{\varepsilon}}{100}$ by (B.9). For the term $\frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2}$, we have $\frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2} \leq \frac{1}{1-\frac{\sqrt{\varepsilon}}{100}}$ by (B.9). Plugging these three inequalities into (B.10), we have

$$\widehat{I}(X_i; X_j \mid X_k) = \frac{1}{2} \log \left(1 + \widehat{\gamma}_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \cdot \frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2} \cdot \frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2} \right) \leq \frac{1}{2} \log \left(1 + \frac{\varepsilon}{100} \cdot \frac{1 + \frac{\sqrt{\varepsilon}}{100}}{1 - \frac{\sqrt{\varepsilon}}{100}} \right) \leq \frac{\varepsilon}{100}$$

for any sufficiently small $\varepsilon > 0$.

We now prove if $I(X_i; X_j \mid X_k) \geq \varepsilon$, then $\widehat{I}(X_i; X_j \mid X_k) > \frac{1}{20} I(X_i; X_j \mid X_k) - \frac{\varepsilon}{40}$. Since $I(X_i; X_j \mid X_k) \geq \varepsilon$, it means that $I(X_i; X_j \mid X_k) = \frac{1}{2} \log(1 + \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2}) \geq \varepsilon$ and hence $\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \geq e^{2\varepsilon} - 1 \geq 2\varepsilon$. We have $\widehat{\gamma}_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \geq \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \sqrt{\frac{\varepsilon}{100}} \geq 0$. For the term $\frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2}$, we have $\frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2} \geq 1 - \frac{\sqrt{\varepsilon}}{100}$ by (B.9). For the term $\frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2}$, we have $\frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2} \geq \frac{1}{1+\frac{\sqrt{\varepsilon}}{100}}$ by (B.9). Plugging these three inequalities into (B.10), we have

$$\widehat{I}(X_i; X_j \mid X_k) = \frac{1}{2} \log \left(1 + \widehat{\gamma}_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \cdot \frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2} \cdot \frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2} \right) \geq \frac{1}{2} \log \left(1 + \left(\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \sqrt{\frac{\varepsilon}{100}} \right)^2 \cdot \frac{1 - \frac{\sqrt{\varepsilon}}{100}}{1 + \frac{\sqrt{\varepsilon}}{100}} \right).$$

Note that, for any a, b, we have $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$ which implies the term $\left(\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \sqrt{\frac{\varepsilon}{100}} \right)^2$ is larger than $\frac{1}{2}\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \frac{\varepsilon}{100}$. Namely, we have

$$\begin{aligned}
\widehat{I}(X_i; X_j \mid X_k) &\geq \frac{1}{2} \log \left(1 + \left(\frac{1}{2} \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \frac{\varepsilon}{100} \right) \cdot \frac{1 - \frac{\sqrt{\varepsilon}}{100}}{1 + \frac{\sqrt{\varepsilon}}{100}} \right) \\
&\geq \frac{1}{2} \log \left(1 + \frac{1}{3} \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \frac{\varepsilon}{100} \right) \\
&\geq \frac{1}{2} \log \left(1 + \frac{1}{3} \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \right) - \frac{\varepsilon}{40}
\end{aligned}$$

for any sufficiently small $\varepsilon > 0$. Note that, for any $a > 0$, $\log(1 + \frac{1}{3}a) \geq \frac{1}{10} \log(1 + a)$. Namely, we have

$$\widehat{I}(X_i; X_j \mid X_k) \geq \frac{1}{2} \log \left(1 + \frac{1}{3} \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \right) - \frac{\varepsilon}{40} \geq \frac{1}{20} \log \left(1 + \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \right) - \frac{\varepsilon}{40} = \frac{1}{20} I(X_i; X_j \mid X_k) - \frac{\varepsilon}{40}. \quad \square$$

B.3 Distribution Learning Upper Bounds

In this subsection, we give the formal proof of the upper bounds on the sample complexity for distribution learning in the non-realizable setting Theorem 3.1 and realizable setting Theorem 3.2:

B.3.1 Non-realizable Case

Theorem 3.1. *Let P be a Gaussian distribution. Given n i.i.d. samples from P, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d^2}{\varepsilon^2} \log \frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 1 satisfies*

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon,$$

with probability at least $1 - \delta$.

Proof. Let T^* be $\arg\min_{T\in\mathcal{T}} D_{\mathrm{KL}}(P \,\|\, P_T)$. By (3.1), we express $D_{\mathrm{KL}}(P \,\|\, P_{\widehat{T}}) - D_{\mathrm{KL}}(P \,\|\, P_{T^*})$ as

$$D_{\mathrm{KL}}(P \,\|\, P_{\widehat{T}}) - D_{\mathrm{KL}}(P \,\|\, P_{T^*}) = -\sum_{(W,Z)\in\widehat{T}} I(W;Z) + \sum_{(X,Y)\in T^*} I(X;Y)$$

Since $\widehat{T}$ is the output of Algorithm 1, we have

$$\sum_{(X,Y)\in T^*} \widehat{I}(X;Y) - \sum_{(W,Z)\in\widehat{T}} \widehat{I}(W;Z) \le 0.$$

Hence, we have

$$\begin{aligned} &D_{\mathrm{KL}}(P \,\|\, P_{\widehat{T}}) - D_{\mathrm{KL}}(P \,\|\, P_{T^*}) \\ &\quad\le \sum_{(W,Z)\in\widehat{T}} \widehat{I}(W;Z) - \sum_{(W,Z)\in\widehat{T}} I(W;Z) + \sum_{(X,Y)\in T^*} I(X;Y) - \sum_{(X,Y)\in T^*} \widehat{I}(X;Y) \end{aligned}$$

By the definition in (B.5) and Corollary B.3, we can show that each $|\widehat{I}(X,Y) - I(X,Y)| < \frac{\varepsilon}{d}$ for all (X,Y) using $O(\frac{d^2}{\varepsilon^2}\log\frac{d}{\delta})$ samples. Therefore, we have

$$D_{\mathrm{KL}}(P \,\|\, P_{\widehat{T}}) - D_{\mathrm{KL}}(P \,\|\, P_{T^*}) < \varepsilon.$$

□

B.3.2 Realizable Case

Fact B.6 ((Bhattacharyya et al., 2023)). *Let T_1 and T_2 be two spanning trees on d vertices such that their symmetric difference consists of the edges $E = \{e_1, e_2, \ldots, e_l\} \in T_1 \setminus T_2$ and $F = \{f_1, f_2, \ldots, f_l\} \in T_2 \setminus T_1$. Then E and F can be paired up, say $\langle e_i, f_i\rangle$, such that for all i, $T_1 \cup \{f_i\} \setminus \{e_i\}$ is a spanning tree.*

Theorem 3.2. *Let T^* be a directed tree and P_{T^*} be a T^*-structured Gaussian. Given n i.i.d. samples from P_{T^*}, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d}{\varepsilon}\log\frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 1 satisfies*

$$D_{\mathrm{KL}}(P_{T^*} \,\|\, P_{\widehat{T}}) \le \varepsilon,$$

with probability at least $1-\delta$.

Proof. We first consider the edge difference between $\widehat{T}$ and T^*. By Fact B.6, we can pair up the edges in $\widehat{T}\setminus T^*$ with the edges in $T^*\setminus\widehat{T}$ such that $T^* \cup \{(W,Z)\}\setminus\{(X,Y)\}$ is also a spanning tree for any $(W,Z) \in \widehat{T}\setminus T^*$ and $(X,Y) \in T^*\setminus\widehat{T}$. Let $\widehat{T}\setminus T^*$ be $\{(W_1,Z_1),\ldots,(W_k,Z_k)\}$ and $T^*\setminus\widehat{T}$ be $\{(X_1,Y_1),\ldots,(X_k,Y_k)\}$ such that (W_i,Z_i) pairs up with (X_i,Y_i) for $i = 1,\ldots,k$. Because of that, there exists a path in T^* from W_i to Z_i containing X_i and Y_i. Without loss of generality, we assume that the order of them is $W_i \rightsquigarrow X_i - Y_i \rightsquigarrow Z_i$ in T^*.

Since $\widehat{T}$ is the output of Algorithm 1, we have

$$\sum_{i=1}^k \widehat{I}(X_i;Y_i) - \sum_{i=1}^k \widehat{I}(W_i;Z_i) \le 0$$

by the definition of the maximal spanning tree. We first expand the LHS as

$$\begin{aligned} \sum_{i=1}^k \widehat{I}(X_i,Y_i) - \sum_{i=1}^k \widehat{I}(W_i,Z_i) &= \sum_{i=1}^k \Big(\widehat{I}(X_i,Y_i) - \widehat{I}(X_i;Z_i) + \widehat{I}(X_i;Z_i) - \widehat{I}(W_i;Z_i)\Big) \\ &= \sum_{i=1}^k \Big(\widehat{I}(X_i;Y_i \mid Z_i) - \widehat{I}(X_i;Z_i \mid Y_i) + \widehat{I}(X_i;Z_i \mid W_i) - \widehat{I}(W_i;Z_i \mid X_i)\Big) \qquad \text{by (B.8)} \\ &= \underbrace{\sum_{i=1}^k \Big(\widehat{I}(X_i;Y_i \mid Z_i) + \widehat{I}(X_i;Z_i \mid W_i)\Big)}_{:=A} - \underbrace{\sum_{i=1}^k \Big(\widehat{I}(X_i;Z_i \mid Y_i) + \widehat{I}(W_i;Z_i \mid X_i)\Big)}_{:=B}. \end{aligned}$$

In other words, we have $A \le B$.

Recall that there exists a path $W_i \rightsquigarrow X_i - Y_i \rightsquigarrow Z_i$ in T^* and hence $(X_i, Z_i) \notin T^*$ which further implies $I(X_i; Z_i \mid Y_i) = 0$. Similarly, we have $I(W_i; Z_i \mid X_i) = 0$. By Theorem B.5 with $\Theta(\frac{1}{\varepsilon'} \log \frac{d}{\delta})$ samples, we have

$$\widehat{I}(X_i; Z_i \mid Y_i) \le \varepsilon'/100 \qquad \text{and} \qquad \widehat{I}(W_i; Z_i \mid X_i) \le \varepsilon'/100 \qquad \text{for all } i = 1, \dots, k.$$

Plugging them into each term in B, we can bound B by $2k \cdot \varepsilon'/100 \le d\varepsilon'/50$. Namely, we have

$$A = \sum_{i=1}^{k} \left(\widehat{I}(X_i; Y_i \mid Z_i) + \widehat{I}(X_i; Z_i \mid W_i) \right) \le d\varepsilon'/50.$$

By Theorem B.5 with $\Theta(\frac{1}{\varepsilon'} \log \frac{d}{\delta})$ samples, we have

$$\frac{1}{20} I(X_i; Y_i \mid Z_i) - \frac{\varepsilon'}{40} \le \widehat{I}(X_i; Y_i \mid Z_i) \qquad \text{and} \qquad \frac{1}{20} I(X_i; Z_i \mid W_i) - \frac{\varepsilon'}{40} \le \widehat{I}(X_i; Z_i \mid W_i)$$

for all $i = 1, \dots, k$. In other words,

$$A \ge \frac{1}{20} \sum_{i=1}^{k} \left(I(X_i; Y_i \mid Z_i) + I(X_i; Z_i \mid W_i) \right) - \frac{d\varepsilon'}{40}$$

or

$$\sum_{i=1}^{k} \left(I(X_i; Y_i \mid Z_i) + I(X_i; Z_i \mid W_i) \right) \le \frac{9d\varepsilon'}{10} \tag{B.11}$$

Now, we can bound $D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}})$. We express it as

$$\begin{aligned} D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}}) &= \sum_{i=1}^{k} I(X_i; Y_i) - \sum_{i=1}^{k} I(W_i; Z_i) = \sum_{i=1}^{k} \left(I(X_i; Y_i) - I(X_i; Z_i) + I(X_i; Z_i) - I(W_i; Z_i) \right) \\ &= \sum_{i=1}^{k} \left(I(X_i; Y_i \mid Z_i) - I(X_i; Z_i \mid Y_i) + I(X_i; Z_i \mid W_i) - I(W_i; Z_i \mid X_i) \right) \qquad \text{by (B.7)} \end{aligned}$$

Recall that we have $I(X_i; Z_i \mid Y_i) = 0$ and $I(W_i; Z_i \mid X_i) = 0$. Combining with (B.11), we have

$$D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}}) \le \frac{9d\varepsilon'}{10}$$

with probability at least $1 - \delta$. By picking $\varepsilon' = \frac{10\varepsilon}{9d}$, we conclude our result. □

B.4 Distribution Learning Lower Bounds

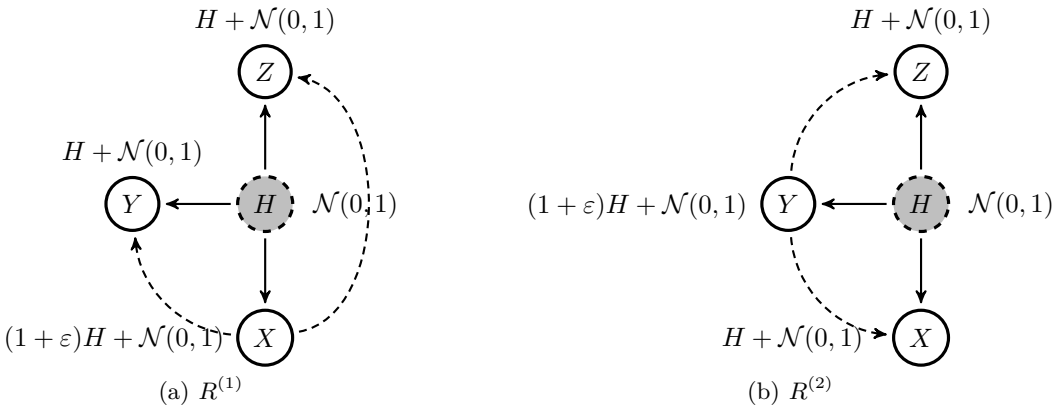
To show the lower bounds, our main idea is to reduce Problem B.7 defined below to our problem.

Problem B.7. *Suppose $R^{(1)}$ and $R^{(2)}$ are two distributions such that $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)}) \le \delta$. Let P be a distribution on m variables where each variable is distributed as either $R^{(1)}$ or $R^{(2)}$ uniformly and independently. We are given n i.i.d. samples drawn from a distribution P. Our task is to determine which distribution the samples are drawn from correctly for at least $51m/100$ variables. Formally, we define*

$$\mathcal{R} := \{(R_1, \dots, R_m) \mid R_i \in \{R^{(1)}, R^{(2)}\}\}.$$

We pick a distribution uniformly from $\mathcal{R}$ and let $P = (R_1^*, \dots, R_m^*)$ be this distribution. Then, our goal is to design an algorithm that takes n i.i.d. samples drawn from P as input and returns $(\widehat{R}_1, \dots, \widehat{R}_m)$ such that $\widehat{R}_i = R_i^*$ for at least $51m/100$ of $\{1, \dots, m\}$.

Fact B.8. *By the standard information-theoretic lower bounds, if $n = o(\frac{1}{\delta})$, then no algorithm can solve Problem B.7 with probability $2/3$.*



(a) $R^{(1)}$ (b) $R^{(2)}$

Figure 4: The $\Omega(1/\varepsilon^2)$ bound in the non-realizable setting. The underlying graph is represented with solid lines, while the best estimated tree structure is depicted with dashed lines.

B.4.1 Non-realizable Case

We define two distributions $Q^{(1)}, Q^{(2)}$ as follows.

$$Q^{(1)} = \begin{cases} H \sim \mathcal{N}(0,1) \\ X \sim (1+\varepsilon)H + \mathcal{N}(0,1) \\ Y \sim H + \mathcal{N}(0,1) \\ Z \sim H + \mathcal{N}(0,1) \end{cases} \quad \text{and} \quad Q^{(2)} = \begin{cases} H \sim \mathcal{N}(0,1) \\ X \sim H + \mathcal{N}(0,1) \\ Y \sim (1+\varepsilon)H + \mathcal{N}(0,1) \\ Z \sim H + \mathcal{N}(0,1) \end{cases} \tag{B.12}$$

Also, we define $R^{(1)}, R^{(2)}$ to be the corresponding marginal distributions on (X, Y, Z).

Lemma B.9. *Suppose R^* is one of $R^{(1)}$ and $R^{(2)}$ defined in (B.12). For any small $\varepsilon > 0$, if a direct tree $\widehat{T}$ satisfies*

$$D_{\mathrm{KL}}(R^* \parallel R^*_{\widehat{T}}) \le \min_T D_{\mathrm{KL}}(R^* \parallel R^*_T) + \frac{\varepsilon}{100} \tag{B.13}$$

and $\widehat{R} = \arg\min_{R \in \{R^{(1)}, R^{(2)}\}} D_{\mathrm{KL}}(R \parallel R_{\widehat{T}})$, then $\widehat{R} = R^*$.

Proof. Since there are three variables, there are only three possible tree structures: $T_1 = Y - X - Z$, $T_2 = X - Y - Z$ and $T_3 = X - Z - Y$. Recall that, by (3.1), we have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) - D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) = I(X;Z) - I(Y;Z) \ge \frac{\varepsilon}{50} \qquad \text{by a straightforward calculation} \tag{B.14}$$

and, similarly, we also have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_3}) - D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) \ge \frac{\varepsilon}{50} \tag{B.15}$$

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_1}) - D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_2}) \ge \frac{\varepsilon}{50} \tag{B.16}$$

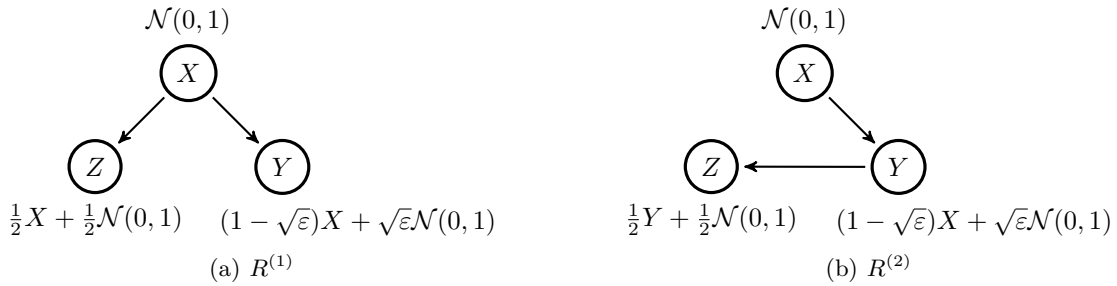
$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_3}) - D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_2}) \ge \frac{\varepsilon}{50} \tag{B.17}$$

By (B.13), (B.15) and (B.17), we have $\widehat{T} \neq T_3$. Namely, $\widehat{T}$ is either T_1 or T_2 (WLOG, say T_1). By (B.13) and (B.16), we have $R^* = R^{(1)}$. By (B.14), we have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) \le D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) - \frac{\varepsilon}{50} < \underbrace{D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) = D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_1})}_{\text{by symmetry}}.$$

Hence, $\widehat{R} = R^{(1)} = R^*$ by the definition of $\widehat{R}$.

□



(a) $R^{(1)}$ (b) $R^{(2)}$

Figure 5: Realizable setting

Theorem 3.3. *Suppose P is an unknown Gaussian distribution. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d^2/\varepsilon^2)$, no algorithm returns a directed tree $\widehat{T}$ such that*

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \le \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon$$

with probability at least $2/3$.

Proof. We will prove the statement by reducing Problem B.7 to our problem. We first split the d variables into $m = d/3$ groups of 3 variables and for each group we select $R^{(1)}$ or $R^{(2)}$ defined in (B.12) (replacing ε with ε/d) uniformly and independently and notice that $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)}) = O(\varepsilon^2/d^2)$ by a straightforward calculation. By Fact B.8, it implies that if $n = o(\frac{d^2}{\varepsilon^2})$ then no algorithm can determine which distribution the samples are drawn from correctly for at least $51m/100$ groups with probability $\frac{2}{3}$.

Suppose there is an algorithm that takes these n i.i.d. samples as input and returns a directed tree $\widehat{T}$ such that

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \le \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon \tag{B.18}$$

with probability $\frac{2}{3}$. If we manage to show that we can use $\widehat{T}$ to determine which distribution the samples are drawn from correctly for $51m/100$ groups then it implies $n = \Omega(\frac{d^2}{\varepsilon^2})$.

We construct the reduction as follows. For the i-th group of variables, we consider its subtree $\widehat{T}_i$ of $\widehat{T}$ and declare $\widehat{R}_i$ to be the distribution for this group where $\widehat{R}_i$ is defined to be $\arg\min_{R \in \{R^{(1)}, R^{(2)}\}} D_{\mathrm{KL}}(R \parallel R_{\widehat{T}_i})$. To see the correctness, we have the following. Since each group is independent, (B.18) can be decomposed into

$$\sum_{i=1}^{m} D_{\mathrm{KL}}(P_i \parallel (P_i)_{\widehat{T}_i}) \le \sum_{i=1}^{m} \min_{T_i} D_{\mathrm{KL}}(P_i \parallel (P_i)_{T_i}) + \varepsilon$$

where P_i is the random pick of $R^{(1)}$ or $R^{(2)}$ for the i-th group. Therefore, at least $51m/100$ of the terms $D_{\mathrm{KL}}(P_i \parallel (P_i)_{\widehat{T}_i}) - \min_{T_i} D_{\mathrm{KL}}(P_i \parallel (P_i)_{T_i}) \le \frac{10\varepsilon}{m}$. By Lemma B.9, for these $51m/100$ groups, $\widehat{R}_i$ is correctly determined, i.e. $\widehat{R}_i = P_i$ and hence the reduction is completed. □

B.4.2 Realizable Case

We define two distributions $R^{(1)}, R^{(2)}$ as follows.

$$R^{(1)} = \begin{cases} X \sim \mathcal{N}(0,1) \\ Y \sim (1-\sqrt{\varepsilon})X + \sqrt{\varepsilon}\mathcal{N}(0,1) \\ Z \sim \frac{1}{2}X + \frac{1}{2}\mathcal{N}(0,1) \end{cases} \quad \text{and} \quad R^{(2)} = \begin{cases} X \sim \mathcal{N}(0,1) \\ Y \sim (1-\sqrt{\varepsilon})X + \sqrt{\varepsilon}\mathcal{N}(0,1) \\ Z \sim \frac{1}{2}Y + \frac{1}{2}\mathcal{N}(0,1) \end{cases} \tag{B.19}$$

Namely, the underlying graph for $R^{(1)}$ is $Y < -X- > Z$ and the underlying graph for $R^{(2)}$ is $X- > Y- > Z$. Both have $X- > Y$ and the only difference is Z.

Lemma B.10. *Suppose R^* is one of $R^{(1)}$ and $R^{(2)}$ defined in (B.19). For any small $\varepsilon > 0$, if a direct tree $\widehat{T}$ satisfies*

$$D_{\mathrm{KL}}(R^* \parallel R^*_{\widehat{T}}) \le \frac{\varepsilon}{100} \tag{B.20}$$

and $\widehat{R} = \arg\min_{R \in \{R^{(1)}, R^{(2)}\}} D_{\mathrm{KL}}(R \parallel R_{\widehat{T}})$, then $\widehat{R} = R^*$.

Proof. Since there are three variables, there are only three possible tree structures: $T_1 = Y - X - Z$, $T_2 = X - Y - Z$ and $T_3 = X - Z - Y$. Recall that, by (3.1), we have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) - D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) = I(X;Z) - I(Y;Z) \ge \frac{\varepsilon}{50} \qquad \text{by a straightforward calculation.}$$

Note that $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) = 0$ and hence

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) \ge \frac{\varepsilon}{50} \tag{B.21}$$

Similarly, we also have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_3}) \ge \Omega(1) \ge \frac{\varepsilon}{50} \tag{B.22}$$

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_1}) \ge \frac{\varepsilon}{50} \tag{B.23}$$

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_3}) \ge \Omega(1) \ge \frac{\varepsilon}{50} \tag{B.24}$$

By (B.20), (B.22) and (B.24), we have $\widehat{T} \neq T_3$. Namely, $\widehat{T}$ is either T_1 or T_2. If $\widehat{T} = T_1$, by (B.20) and (B.23), we have

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{\widehat{T}}) > D_{\mathrm{KL}}(R^* \parallel R^*_{\widehat{T}})$$

and hence $R^* = R^{(1)}$. If $\widehat{T} = T_2$, by (B.20) and (B.21), we have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{\widehat{T}}) > D_{\mathrm{KL}}(R^* \parallel R^*_{\widehat{T}})$$

and hence $R^* = R^{(2)}$. By the definition of $\widehat{R}$, both cases imply $\widehat{R} = R^*$.

□

Theorem 3.4. *Suppose P is an unknown Gaussian distribution such that there exists a directed tree T^* that P is T^*-structured, i.e. $P = P_{T^*}$. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d/\varepsilon)$, no algorithm returns a directed tree $\widehat{T}$ such that*

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \le \varepsilon$$

with probability at least $2/3$.

Proof. We will prove the statement by reducing Problem B.7 to our problem. We first split the d variables into $m = d/3$ groups of 3 variables and for each group we select $R^{(1)}$ or $R^{(2)}$ defined in (B.19) (replacing ε with ε/d) uniformly and independently and notice that $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)}) = O(\varepsilon/d)$ by a straightforward calculation. By Fact B.8, it implies that if $n = o(\frac{d}{\varepsilon})$ then no algorithm can determine which distribution the samples are drawn from correctly for at least $51m/100$ groups with probability $\frac{2}{3}$.

Suppose there is an algorithm that takes these n i.i.d. samples as input and returns a directed tree $\widehat{T}$ such that

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \le \varepsilon \tag{B.25}$$

with probability $\frac{2}{3}$. If we manage to show that we can use $\widehat{T}$ to determine which distribution the samples are drawn from correctly for $51m/100$ groups then it implies $n = \Omega(\frac{d}{\varepsilon})$.

Algorithm 3: ORIENT algorithm

Input: Skeleton $\widehat{T}$, separation sets S

Output: CPDAG $\widehat{\widehat{T}}$.

1. For all pairs of nonadjacent nodes j, k with common neighbour ℓ:
   (a) If $\ell \notin S(j,k)$, then directize $j - \ell - k$ in $\widehat{T}$ by $j \to \ell \leftarrow k$
2. In the resulting PDAG $\widehat{T}$, orient as many as possible undirected edges by applying following rules:
   - **R1** Orient $k - \ell$ into $k \to \ell$ whenever there is an arrow $j \to k$ such that j and ℓ are not adjacent
   - **R2** Orient $j - k$ into $j \to k$ whenever there is a chain $j \to \ell \to k$
   - **R3** Orient $j - k$ into $j \to k$ whenever there are two chains $j - \ell \to k$ and $j - i \to k$ such that ℓ and i are not adjacent
   - **R4** Orient $j - k$ into $j \to k$ whenever there are two chains $j - \ell \to i$ and $\ell - i \to k$ such that ℓ and i are not adjacent
3. Return $\widehat{T}$ as $\widehat{\widehat{T}}$.

We construct the reduction as follows. For the i-th group of variables, we consider its subtree $\widehat{T}_i$ of $\widehat{T}$ and declare $\widehat{R}_i$ to be the distribution for this group where $\widehat{R}_i$ is defined to be $\arg\min_{R \in \{R^{(1)}, R^{(2)}\}} D_{\mathrm{KL}}(R \parallel R_{\widehat{T}_i})$. To see the correctness, we have the following. Since each group is independent, (B.25) can be decomposed into

$$\sum_{i=1}^{m} D_{\mathrm{KL}}(P_i \parallel (P_i)_{\widehat{T}_i}) \le \varepsilon$$

where P_i is the random pick of $R^{(1)}$ or $R^{(2)}$ for the i-th group. Therefore, at least $51m/100$ of the terms $D_{\mathrm{KL}}(P_i \parallel (P_i)_{\widehat{T}_i}) \le \frac{10\varepsilon}{m}$. By Lemma B.10, for these $51m/100$ groups, $\widehat{R}_i$ is correctly determined, i.e. $\widehat{R}_i = P_i$ and hence the reduction is completed.

□

B.5 Learning Polytrees given Skeleton

In this section, we sketch how to obtain a sample-efficient algorithm for learning bounded-degree gaussian *polytrees* by adapting the recent results from (Choo et al., 2023), using the guarantees of the estimator $\widehat{I}$, assuming that the skeleton is known. Let a *m-polytree* denote a polytree with maximum in-degree m. Our main result in this section is the following:

Theorem B.11. *There exists an algorithm which, given n samples from a gaussian m-polytree P over $\mathbb{R}^d$, accuracy parameter $\varepsilon > 0$, failure probability δ, maximum in-degree m, and the explicit description of the ground truth skeleton of P, outputs a m-polytree $\widehat{P}$ such that $D_{\mathrm{KL}}(P \| \widehat{P}) \le \varepsilon$ with success probability at least $1 - \delta$, as long as:*

$$n \ge \widetilde{O}\left(\frac{d}{\varepsilon} \log \frac{1}{\delta}\right).$$

Moreover, the algorithm runs in time polynomial in n and d.

Note that the guarantee in Theorem B.11 is entirely independent of any faithfulness parameter, in contrast to Theorem 4.3. The algorithm and its analysis is exactly the same as in Choo et al. (2023), with the only change being that we use (B.5) for the estimator $\widehat{I}$.

C PROOFS OF Section 4

C.1 Sample Conditional Correlation Coefficient as CI Tester

`PC-Tree` relies on sample (conditional) correlation coefficient as (conditional) independence tester. Specifically, denote the sample covariance matrix to be $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)} X^{(i)\top}$, for any two nodes $j, k \in V$ and any subset $S \subseteq V \setminus \{j,k\}$, which could be $\emptyset$, the sample correlation coefficient is defined by

$$\widehat{\rho}_{jk\,|\,S} := \frac{\widehat{\Sigma}_{jk} - \widehat{\Sigma}_{jS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sk}}{\sqrt{(\widehat{\Sigma}_{jj} - \widehat{\Sigma}_{jS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sj})(\widehat{\Sigma}_{kk} - \widehat{\Sigma}_{kS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sk})}}$$

Then the conditional independence tester for hypothesis $H_0 : X_j \perp\!\!\!\perp X_k \,|\, X_S$ is given by a cutoff on the sample correlation coefficient:

$$\text{Output} = \begin{cases} \text{accept } H_0 & \text{if } |\widehat{\rho}_{jk\,|\,S}| \geq c/2 \\ \text{reject } H_0 & \text{if } |\widehat{\rho}_{jk\,|\,S}| < c/2 \end{cases}. \tag{C.1}$$

Here the choice of $c/2$ is for theoretical purpose. Since correlation coefficient is normalized between $[-1,1]$, in practice, the tester can be implemented by choosing a cutoff that is small enough, e.g. 0.05. The analysis of `PC-Tree` crucially relies on the following lemma on the estimation error of the sample (conditional) correlation coefficients:

Lemma C.1. *Let $X \in \mathbb{R}^d \sim \mathcal{N}(0,\Sigma)$, for any $j,k \in V$ and any subset $S \subseteq V \setminus \{j,k\}$ with $|S| \leq q$, $\delta \in (0,1)$, if $n \gtrsim q + 1/\delta^2$, then*

$$\Pr(|\widehat{\rho}_{jk\,|\,S} - \rho_{jk\,|\,S}| \geq \delta) \leq \exp(-C_0(n-q)\delta^2)\,,$$

for some universal constant $C_0 > 0$.

It is clear to see that as long as the (conditional) correlation coefficients are estimated accurately enough, the CI tests are correct due to c-strong Tree-faithfulness. Lemma C.1 is more general than needed to analyze `PC-Tree` algorithm. Since Lemma C.1 reveals the dependence on the size of conditioning set S, while `PC-Tree` only requires $|S| \leq 1$.

C.2 Proof of Lemma C.1

Lemma C.1. *Let $X \in \mathbb{R}^d \sim \mathcal{N}(0,\Sigma)$, for any $j,k \in V$ and any subset $S \subseteq V \setminus \{j,k\}$ with $|S| \leq q$, $\delta \in (0,1)$, if $n \gtrsim q + 1/\delta^2$, then*

$$\Pr(|\widehat{\rho}_{jk\,|\,S} - \rho_{jk\,|\,S}| \geq \delta) \leq \exp(-C_0(n-q)\delta^2)\,,$$

for some universal constant $C_0 > 0$.

Proof. The proof is a combination of the following lemmas. We start with analyzing sample marginal correlation of bivariate normal distribution, then extend to conditional correlation.

Lemma C.2. *Let $W = (X,Y) \sim \mathcal{N}(0,\Sigma)$ where $\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$, and $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$. Let the sample covariance matrix and correlation be*

$$\frac{1}{n}\sum_{\ell=1}^{n} w^{(\ell)} w^{(\ell)\top} = \begin{pmatrix} \widehat{\sigma}_X^2 & \widehat{\sigma}_{XY} \\ \widehat{\sigma}_{XY} & \widehat{\sigma}_Y^2 \end{pmatrix}, \quad \text{and} \quad \widehat{\rho} = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}_X \widehat{\sigma}_Y}.$$

For $\delta \in (0,1)$, if $n \gtrsim 1/\delta^2$, then

$$\Pr(|\widehat{\rho} - \rho| \geq \delta) \leq \exp(-C_0 n \delta^2)\,,$$

for some constant $C_0 > 0$.

Now look at sample conditional correlation, suppose we want to estimate $\rho_{jk\,|\,S}$ with $|S| = q' \le q$. Recall the sample covariance matrix is $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X^{(i)} X^{(i)\top}$. Denote $I = \{j,k\}$, then the estimator is given by 2×2 matrix

$$\widehat{\Sigma}_{II\,|\,S} := \widehat{\Sigma}_{II} - \widehat{\Sigma}_{II,S}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{S,II}\,.$$

We borrow a classic result regarding the distribution of $\widehat{\Sigma}_{II\,|\,S}$:

Lemma C.3 (Anderson (1958), Theorem 4.3.4)**.** *The sample covariance matrix $\widehat{\Sigma}_{II\,|\,S}$ is distributed as $\frac{1}{n}\sum_{\ell=1}^{n-q'} u^{(\ell)}u^{(\ell)\top}$, where $\{u^{(\ell)}\}_{\ell=1}^{n-q'}$ are independently distributed according to $\mathcal{N}(0,\Sigma_{II\,|\,S})$.*

Then applying the bivariate result from Lemma C.2 with covariance matrix $\Sigma_{II\,|\,S}$ and sample size $n-q' \le n-q$ completes the proof. □

It remains to prove the lemma used in proof above.

Proof of Lemma C.2. Let $Z_X = X/\sigma_X$, $Z_Y = Y/\sigma_Y$, then $Z_X, Z_Y \sim \mathcal{N}(0,1)$ and $\rho_{Z_X,Z_Y} = \rho = \mathrm{cov}(Z_X,Z_Y) \in [-1,1]$. Denote the corresponding samples to be $z_X = (z_X^{(1)},\dots,z_X^{(n)})$ and $z_Y = (z_Y^{(1)},\dots,z_Y^{(n)})$, therefore

$$\widehat{\rho} = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}_X\widehat{\sigma}_Y} = \frac{\widehat{\sigma}_{XY}/(\sigma_X\sigma_Y)}{(\widehat{\sigma}_X/\sigma_X)\times(\widehat{\sigma}_Y/\sigma_Y)} = \frac{\langle z_X, z_Y\rangle}{\|z_X\|\|z_Y\|}\,.$$

Then the deviation

$$\begin{aligned}
|\widehat{\rho}-\rho| &= \left|\frac{\langle z_X,z_Y\rangle}{\|z_X\|\|z_Y\|} - \mathrm{cov}(Z_X,Z_Y)\right| \\
&\le \left|\frac{\langle z_X,z_Y\rangle/n}{\|z_X\|\|z_Y\|/n} - \frac{\mathrm{cov}(Z_X,Z_Y)}{\|z_X\|\|z_Y\|/n} + \frac{\mathrm{cov}(Z_X,Z_Y)}{\|z_X\|\|z_Y\|/n} - \mathrm{cov}(Z_X,Z_Y)\right| \\
&\le \left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right|\left|\langle z_X,z_Y\rangle/n - \mathrm{cov}(Z_X,Z_Y)\right| + \left|\langle z_X,z_Y\rangle/n - \mathrm{cov}(Z_X,Z_Y)\right| \\
&\quad + \left|\mathrm{cov}(Z_X,Z_Y)\right|\left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right|.
\end{aligned}$$

We apply the following lemma to bound the errors:

Lemma C.4. *If $(X,Y) \sim \mathcal{N}\Big(0, \begin{pmatrix} 1 & r \\ r & 1\end{pmatrix}\Big)$ for $|r|\le 1$, then the sample variance $\widehat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n {X^{(i)}}^2$, $\widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^n {Y^{(i)}}^2$ and sample covariance $\widehat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^n X^{(i)}Y^{(i)}$ have the following bounds: for $\zeta<1$, if $n \ge \frac{2048\log 7}{\zeta^2}$, then*

$$\begin{aligned}
&\Pr(|\widehat{\sigma}_X^2 - 1| \ge \zeta) \le \exp(-n\zeta^2/16) \\
&\Pr(|\widehat{\sigma}_Y^2 - 1| \ge \zeta) \le \exp(-n\zeta^2/16) \\
&\Pr(|\widehat{\sigma}_{XY} - r| \ge \zeta) \le \exp(-n\zeta^2/2048)\,.
\end{aligned}$$

Using Lemma C.4, with probability at least $1-3\exp(-n\zeta^2/2048)$, we have $|\|z_X\|^2/n - 1| \le \zeta$, $|\|z_Y\|^2/n - 1| \le \zeta$, $|\langle z_X,z_Y\rangle/n - \mathrm{cov}(Z_X,Z_Y)| \le \zeta$. Then

$$\left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right| = \frac{|\|z_X\|\|z_Y\|/n - 1|}{\|z_X\|\|z_Y\|/n} \le \frac{\zeta}{1-\zeta}\,.$$

Choose $\zeta = \frac{\delta}{3+\delta}$, then $\left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right| \le \delta/3$, $\left|\langle z_X,z_Y\rangle/n - \mathrm{cov}(Z_X,Z_Y)\right| \le \delta/(3+\delta) \le \delta/3$. Lastly,

$$|\widehat{\rho}-\rho| \le \frac{\delta}{3}\times\frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} \le \delta\,,$$

with probability at least

$$\begin{aligned}1-3\exp(-n\zeta^2/2048) &= 1-\exp\Big(-n\times\frac{\delta^2}{(3+\delta)^2}/2048+\log 3\Big)\\ &\geq 1-\exp\Big(-n\times\frac{\delta^2}{16\times 2048}+\log 3\Big)\\ &\geq 1-\exp(-C_0 n\delta^2)\,,\end{aligned}$$

for some constant $C_0>0$ as long as $n\gtrsim 1/\delta^2$. □

Proof of Lemma C.4. We only show variance bound for X. Since $\widehat{\sigma}_X^2\sim\chi_n^2/n$, using the concentration of χ^2 distribution, we have

$$\Pr(|\widehat{\sigma}_X^2-1|\geq\zeta)=\Pr(|\chi_n^2-n|/n\geq\zeta)\leq\exp(-n\zeta^2/16)\,.$$

Now we show bound for covariance. Since bivariate Gaussian (X,Y) can be reparameterized by

$$\begin{aligned}X&=U+W\\ Y&=V+W\end{aligned}$$

where U,V,W are mutually independent with $\mathrm{var}(U)=\mathrm{var}(V)=1-r$, $\mathrm{var}(W)=r$. Therefore,

$$\begin{aligned}\widehat{\sigma}_{XY} &= \frac{1}{n}\sum_{i=1}^n U^{(i)}V^{(i)}+\frac{1}{n}\sum_{i=1}^n U^{(i)}W^{(i)}+\frac{1}{n}\sum_{i=1}^n V^{(i)}W^{(i)}+\frac{1}{n}\sum_{i=1}^n {W^{(i)}}^2\\ &= \frac{1-r}{2n}\Big[\sum_{i=1}^n\Big(\frac{U'^{(i)}+V'^{(i)}}{\sqrt{2}}\Big)^2-\sum_{i=1}^n\Big(\frac{U'^{(i)}-V'^{(i)}}{\sqrt{2}}\Big)^2\Big]\\ &\quad+\frac{\sqrt{r(1-r)}}{2n}\Big[\sum_{i=1}^n\Big(\frac{U'^{(i)}+W'^{(i)}}{\sqrt{2}}\Big)^2-\sum_{i=1}^n\Big(\frac{U'^{(i)}-W'^{(i)}}{\sqrt{2}}\Big)^2\Big]\\ &\quad+\frac{\sqrt{r(1-r)}}{2n}\Big[\sum_{i=1}^n\Big(\frac{V'^{(i)}+W'^{(i)}}{\sqrt{2}}\Big)^2-\sum_{i=1}^n\Big(\frac{V'^{(i)}-W'^{(i)}}{\sqrt{2}}\Big)^2\Big]+\frac{r}{n}\sum_{i=1}^n {W'^{(i)}}^2\\ &\overset{\mathcal{D}}{\sim} \frac{1-r}{2n}(\chi_{n11}^2-\chi_{n12}^2)+\frac{\sqrt{r(1-r)}}{2n}(\chi_{n21}^2-\chi_{n22}^2)+\frac{\sqrt{r(1-r)}}{2n}(\chi_{n31}^2-\chi_{n32}^2)+\frac{r}{n}\chi_{n4}^2\end{aligned}$$

where U',V',W' are standard normal random variables, thus $\sum_{i=1}^n(U'^{(i)}\pm V'^{(i)})^2/2$, $\sum_{i=1}^n(U'^{(i)}\pm W'^{(i)})^2/2$, $\sum_{i=1}^n(V'^{(i)}\pm W'^{(i)})^2/2$ are χ_n^2 random variables. Since $r\leq 1$,

$$\begin{aligned}\Pr(|\widehat{\sigma}_{XY}-r|\geq\zeta) &\leq \Pr\Big(\frac{1-r}{2}\times\frac{1}{n}|\chi_{n11}^2-\chi_{n12}^2|\geq\zeta/4\Big)\\ &\quad+\Pr\Big(\frac{\sqrt{r(1-r)}}{2}\times\frac{1}{n}|\chi_{n21}^2-\chi_{n22}^2|\geq\zeta/4\Big)\\ &\quad+\Pr\Big(\frac{\sqrt{r(1-r)}}{2}\times\frac{1}{n}|\chi_{n31}^2-\chi_{n32}^2|\geq\zeta/4\Big)\\ &\quad+\Pr\Big(r\times|\chi_{n41}^2/n-1|\geq\zeta/4\Big)\\ &\leq \Pr\Big(|\chi_{n11}^2/n-1|\geq\zeta/8\Big)+\Pr\Big(|\chi_{n12}^2/n-1|\geq\zeta/8\Big)\\ &\quad+\Pr\Big(|\chi_{n21}^2/n-1|\geq\zeta/8\Big)+\Pr\Big(|\chi_{n22}^2/n-1|\geq\zeta/8\Big)\\ &\quad+\Pr\Big(|\chi_{n31}^2/n-1|\geq\zeta/8\Big)+\Pr\Big(|\chi_{n32}^2/n-1|\geq\zeta/8\Big)\\ &\quad+\Pr\Big(|\chi_{n41}^2/n-1|\geq\zeta/4\Big)\\ &\leq 7\exp(-n\zeta^2/32^2)\leq\exp(-n\zeta^2/2048)\,.\end{aligned}$$

The last inequality holds when $n\geq 2048\log 7/\zeta^2$. □

C.3 Proof of Theorem 4.3

Theorem 4.3. *For any $T \in \widetilde{\mathcal{T}}$, assuming P is c-strong tree-faithful to T, applying Algorithm 2 with sample correlation for CI testing, if the sample size*

$$n \gtrsim \frac{1}{c^2}\left(\log d + \log(1/\delta)\right),$$

then $\Pr(\widehat{T} = \mathrm{sk}(T)) \geq 1 - \delta$, *and* $\Pr(\textsc{Orient}(\widehat{T}, S) = \overline{T}) \geq 1 - \delta$

Proof. We firstly show the correctness of Algorithm 2. We make following notation of sets of nodes:

- $W = \{(j,k) : 1 \leq j < k \leq d\}$ is the set of all pairs of nodes in $[d]$;
- E is the true edge set;
- $A = \{(j,k) : j \text{ and } k \text{ are d-separated by } \emptyset\}$;
- $B = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j \text{ and } k \text{ are d-separated by } \ell\}$
- $C = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j \to \ell \leftarrow k \text{ is a } v\text{-structure}\}$
- $D = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j - \ell - k \text{ is a unshielded triple but not a } v\text{-structure}\}$

We claim that

1. E and $A \cup B$ are disjoint;
2. $W = E \cup A \cup B$;
3. $C \subseteq A$;
4. $D \subseteq B$.

It is easy to see the first claim, since for any pair of nodes connected by an edge, they cannot be d-separated by any set, and vice versa.

For the second claim, it suffices to show that for any pair of nodes not adjacent, it is in either A or B. First of all, for any two nodes j and k not adjacent, there will be one and only one path, denoted as ϕ, with length at least two between them. By property of polytree:

- If there is a collider on ϕ, then the path is blocked by $\emptyset$, so $(j,k) \in A$;
- If there is no collider on ϕ, then any node on ϕ will block the path, thus there exists $\ell \in [d] \setminus \{j,k\}$ such that i and j are d-separated by ℓ, so $(j,k) \in B$.

For the third claim, since $j \to \ell \leftarrow k$ is the only path between (j,k), which is blocked by $\emptyset$, thus $C \subseteq A$. For the forth claim, since $j - \ell - k$ is the only path between (j,k), either one of $j \to \ell \to k$ and $j \leftarrow \ell \leftarrow k$ and $j \leftarrow \ell \to k$ will be blocked by ℓ, thus $D \subseteq B$.

We now claim if the CI tests in Step 2 of Algorithm 2 are correct for

- all pairs $(j,k) \in E$ with $\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\}$;
- all pairs $(j,k) \in A$ with $\ell = \emptyset$;
- all pairs $(j,k) \in C$ with ℓ being the collider;
- all pairs $(j,k) \in B$ with ℓ being the corresponding separation node(s),

then

1. the returned $\widehat{T}$ has the correct edge set E thus is the correct skeleton;
2. for any $(j,k) \in C$, $\ell \notin S(j,k)$;
3. for any $(j,k) \in D$, $\ell \in S(j,k)$.

For the first claim, if the CI tests conducted in Step 2 are correct for E, then pairs in E will pass all the CI tests and be included into $\widehat{E}$ (which is ensured by adjacency-faithfulness in Tree-faithfulness). But pairs in A will not pass marginal independence tests, and pairs in B will not pass some CI tests with corresponding ℓ (which is ensured by Markov property). Therefore, the returned $\widehat{T}$ is the correct skeleton. The second claim is ensured by orientation-faithfulness in Tree-faithfulness, and the third claim is ensured by Markov property and $D \subseteq B$.

Once the returned $\widehat{T}$ is the correct skeleton, Algorithm 3 will use the returned separation sets to determine v-structure for each possible unshielded triple. Note that $\{\text{All unshielded triples}\} = C \cup D$. For any $(j,k) \in C$, $\ell \notin S(j,k)$, thus it will be oriented as a v-structure; For any $(j,k) \in D$, $\ell \in S(j,k)$; thus it will remain as non-v-structure. Then ORIENT step is correct, which leads to correct CPDAG.

Finally we show the sample complexity of Algorithm 2 with CI tester (C.1). Note that correct CI tests implies correct estimation. Therefore,

$$
\begin{aligned}
&\Pr(\widehat{T} \neq \mathrm{sk}(T)) \\
&\le \Pr\Bigg(\bigcup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{or } \bigcup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{or } \bigcup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{or } \bigcup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} |\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \ge c/2 \Bigg) \\
&\le \binom{d}{2} \times (1 + (d-2)) \times \sup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{or } \sup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{or } \sup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{or } \sup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} \Pr(|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \ge c/2) \\
&\le \exp(3\log d) \times \sup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{or } \sup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{or } \sup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{or } \sup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} \Pr(|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \ge c/2) \\
&\le \exp\Big(-C_0(n-1)c^2 + 3\log d \Big).
\end{aligned}
$$

The first inequality is because it suffices to have $|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \le c/2$ for correct CI test. By c-strong Tree-faithfulness, $|\rho_{ij\,|\,S}| \ge c$ for $\rho_{ij\,|\,S} \neq 0$. Therefore,

$$
\begin{cases}
\widehat{\rho}_{ij\,|\,S} > c/2 & \text{if } \rho_{ij\,|\,S} \neq 0 \\
\widehat{\rho}_{ij\,|\,S} \le c/2 & \text{if } \rho_{ij\,|\,S} = 0
\end{cases}
$$

Thus the cutoff $= c/2$ implies correct CI tests. The last inequality is by Lemma C.1 where $q = 1$ and the sample size requirement is satisfied by the stated sample complexity. Set RHS to be smaller than δ, we need sample complexity

$$
n \gtrsim \frac{1}{c^2}\left(\log d + \log\frac{1}{\delta}\right),
$$

which completes the proof. □

C.4 Proof of Theorem 4.4

Theorem 4.4. *Assuming c-strong tree-faithfulness, and $c^2 \le 1/5$, $d \ge 4$, if the sample size is bounded as*

$$
n \le \frac{1-2\delta}{8} \times \frac{\log d}{c^2},
$$

then for any estimator $\widehat{T}$ for $\overline{T}$,

$$
\inf_{\widehat{T}} \sup_{\substack{T\in\widetilde{\mathcal{T}} \\ P \text{ is } c\text{-strong} \\ \text{tree-faithful to } T}} \Pr(\widehat{T} \neq \overline{T}) \ge \delta - \frac{\log 2}{\log d}.
$$

Proof. We construct a hard ensemble to show the lower bound. The construction is as follows: consider a subset $\mathcal{T}' \subset \mathcal{T} \subset \widetilde{\mathcal{T}}$, where $\mathcal{T}'$ is all the directed trees rooted at the first node $k = 1$. $\mathcal{T}'$ has the same cardinality as all undirected trees with d nodes, and the elements in it have different skeletons and no v-structures. Since our target is MEC, which is determined by its skeleton and v-structures, we have at least as many MECs as undirected trees, which leads to cardinality $|\mathcal{T}'| = d^{d-2}$ using Cayley's formula. Thus the size of the ensemble is lower bounded as

$$\log |\mathcal{T}'| = (d-2)\log d \geq \frac{1}{2} d \log d$$

The inequality holds when d is large enough, e.g. $d \geq 4$. Any directed tree has an important property: each node has at most one parent. Then we parameterize $\mathcal{T}'$ as follows

$$X_k = \beta X_{\mathrm{pa}(k)} + \eta_k\,, \quad \forall k \in [d] \tag{C.2}$$

where $\eta_k \sim \mathcal{N}(0,1)$ for all $k \in [d]$. Now we determine $\beta > 0$ to make sure the parametrization satisfies c-strong Tree-faithfulness.

In the subsequent lemma, we assert that the condition $\beta^2 = 2c^2 \asymp c^2$ is adequate for the validity of c-strong Tree-faithfulness, provided that c is sufficiently small:

Lemma C.5. *If $\beta = \sqrt{2}c$ and $c^2 \leq 1/5$, then for any $T \in \mathcal{T}'$, the distribution defined in (C.2):*

1. *is c-strong Tree-faithful to T;*
2. *for all $k \in [d]$, $\mathrm{var}(X_k) \leq 1 + \frac{\beta^2}{1-\beta^2}$.*

It remains to bound the KL divergence between any two instances in this ensemble. Before that, we claim that for any instance, we have $\mathrm{cov}(X_k, X_j) > 0$ for all distinct $j, k \in [d]$. This is because for any pair of distinct nodes (j,k), there can be 3 possible paths between them:

- There is a directed path $j \to \phi_1 \to \cdots \to \phi_h \to k$ with length $h+1$, then $\mathrm{cov}(X_j, X_k) = \mathbb{E}[X_j X_k] = \beta^{h+1}\,\mathbb{E}[X_j^2] > 0$;
- There is a directed path $k \to \phi_1 \to \cdots \to \phi_h \to j$ with length $h+1$, then $\mathrm{cov}(X_j, X_k) = \mathbb{E}[X_j X_k] = \beta^{h+1}\,\mathbb{E}[X_k^2] > 0$;
- j, k share a common ancestor ℓ and there is a path $j \leftarrow \phi_1 \leftarrow \cdots \leftarrow \phi_h \leftarrow \ell \to \varphi_1 \to \cdots \to \varphi_g \to k$, then $\mathrm{cov}(X_j, X_k) = \mathbb{E}[X_j X_k] = \beta^{h+g+2}\,\mathbb{E}[X_\ell^2] > 0$.

To compute the KL divergence between distributions P_0 and P_1 induced by any two $T_0, T_1 \in \mathcal{T}'$, let's first look at the covariance matrices of them Σ_0, Σ_1. Under our parametrization, they share the same determinant. To see this, let covariance matrix of η be $\Sigma_\eta = I_d$, for $\ell \in \{0,1\}$, $\det(\Sigma_\ell) = \det(\Sigma_\eta) = \det(I_d) = 1$. Then the KL divergence is:

$$\begin{aligned}
D_{\mathrm{KL}}(P_0 \| P_1) &= \mathop{\mathbb{E}}_{P_0} \log \frac{P_0}{P_1} \\
&= \mathop{\mathbb{E}}_{P_0} \log \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^d (X_k - \beta\,\mathrm{pa}_{T_0}(k))^2 \right) / \sqrt{\det(\Sigma_0)}}{\exp\left(-\frac{1}{2}\sum_{k=1}^d (X_k - \beta\,\mathrm{pa}_{T_1}(k))^2 \right) / \sqrt{\det(\Sigma_1)}} \\
&= \frac{1}{2}\left[\mathop{\mathbb{E}}_{P_0} \sum_{k=1}^d (X_k - \beta\,\mathrm{pa}_{T_1}(k))^2 - d \right].
\end{aligned}$$

For all $k \in [d]$, let $\mathrm{pa}_{T_1}(k) = j$, then

$$\begin{aligned}
\mathop{\mathbb{E}}_{P_0} (X_k - \beta\,\mathrm{pa}_{T_1}(k))^2 &= \mathop{\mathbb{E}}_{P_0}[X_k^2] + \beta^2 \mathop{\mathbb{E}}_{P_0}[X_j^2] - 2\beta \mathop{\mathbb{E}}_{P_0}[X_k X_j] \\
&\leq \mathop{\mathbb{E}}_{P_0}[X_k^2] + \beta^2 \mathop{\mathbb{E}}_{P_0}[X_j^2]
\end{aligned}$$

$$\leq (1+\beta^2)\left(1+\frac{\beta^2}{1-\beta^2}\right)$$
$$= 1+\frac{2\beta^2}{1-\beta^2}\,.$$

The first inequality is because all covariances are positive; the second one is due to the upper bound for all variances. Thus, we have

$$D_{\mathrm{KL}}(P_0\|P_1) \leq \frac{1}{2}\left(d+\frac{2d\beta^2}{1-\beta^2}-d\right) = d\beta^2 \times \frac{1}{1-\beta^2} \leq 2d\beta^2 = 4dc^2$$

The last inequality holds when β^2 is small enough, e.g. $\beta^2 \leq 1/2$. The proof follows from applying Fano's inequality with KL divergence upper bound $4dc^2$ and cardinality of ensemble lower bound $\frac{1}{2}d\log d$. □

We end by proving the lemma used in the lower bound proof.

Proof of Lemma C.5. Since for any $T \in \mathcal{T}'$, there is no v-structure because each node has at most one parent, thus it suffices to show the first part of Definition 4.2.

We first show all marginal variances are bounded, i.e. $1 \leq \mathrm{var}(X_k) \leq 1+\beta^2/(1-\beta^2)$ for all $k \in [d]$. Starting from the root node r, whose variance is $\mathrm{var}(X_r) = \mathrm{var}(\eta_r) = 1$, we can compute the variances of its children, they are all $\mathrm{var}(X_\ell) = \mathrm{var}(\eta_\ell) + \beta^2\,\mathrm{var}(X_r) = 1+\beta^2$ for all $\ell \in \mathrm{ch}(r)$. Proceed the calculation, $\mathrm{var}(X_j) = \mathrm{var}(\eta_j) + \beta^2\,\mathrm{var}(X_\ell) = 1+\beta^2+\beta^4$ for all $j \in \mathrm{ch}(\ell)$ and $\ell \in \mathrm{ch}(r)$. Therefore, because the longest path has length at most $d-1$,

$$1 \leq \mathrm{var}(X_k) \leq 1+\beta^2+\beta^4+\cdots+\beta^{2d} = 1+\frac{\beta^2}{1-\beta^2}\times(1-\beta^{2(d-1)}) \leq 1+\frac{\beta^2}{1-\beta^2}\,,\quad \forall k \in [d]$$

Now we can show the marginal correlation is lower bounded for any adjacent nodes (j,k). Without loss of generality, let $j = \mathrm{pa}(k)$, then $X_k = \beta X_j + \eta_k$, and the correlation

$$\rho(X_j, X_k) = \frac{\mathbb{E}[X_jX_k]}{\sqrt{\mathrm{var}(X_k)\,\mathrm{var}(X_j)}} = \beta\sqrt{\frac{\mathbb{E}[X_j^2]}{1+\beta^2\,\mathbb{E}[X_j^2]}}$$

Thus $\rho(X_j,X_k) \geq c \Leftrightarrow \beta^2\,\mathbb{E}[X_j^2] \geq \frac{c^2}{1-c^2}$. Since $\mathbb{E}(X_j^2) \geq 1$, then $\beta^2\,\mathbb{E}(X_j^2) \geq \beta^2 = 2c^2 \geq \frac{c^2}{1-c^2}$ when $c^2 \leq 1/2$. Now consider any pair of adjacent nodes (j,k), assuming $j = \mathrm{pa}(k)$, and any other node $\ell \in [d]\setminus\{j,k\}$, there are 4 cases on the relation between ℓ and (j,k):

1. ℓ is ancestor of j, i.e. a directed path ϕ: $\ell \to \phi_1 \to \cdots \to \phi_h \to j$;
2. j and ℓ share the same ancestor w, i.e. a directed path ϕ: $w \to \phi_1 \to \cdots \to \phi_h \to j$ and a directed path φ: $w \to \varphi_1 \to \cdots \to \varphi_g \to \ell$;
3. ℓ is a descendant of k, i.e. a directed path ϕ: $j \to k \to \phi_1 \to \cdots \to \phi_h \to \ell$;
4. ℓ is a descendant of j but not k, i.e. a directed path ϕ: $j \to \phi_1 \to \cdots \to \phi_h \to \ell$ not going through k;

where $h \geq 0$ in either case. We deal with them separately:

- For the first and second case, because $X_\ell \perp\!\!\!\perp \eta_k$, the conditional correlation is

$$\rho(X_k, X_j \,|\, X_\ell) = \frac{\mathbb{E}[X_kX_j \,|\, X_\ell]}{\sqrt{\mathbb{E}(X_k^2 \,|\, X_\ell)\,\mathbb{E}(X_j^2 \,|\, X_\ell)}}$$
$$= \frac{\beta\,\mathbb{E}[X_j^2 \,|\, X_\ell]}{\sqrt{\mathbb{E}(X_j^2 \,|\, X_\ell)(1+\beta^2\,\mathbb{E}(X_j^2 \,|\, X_\ell))}}$$

$$= \sqrt{\frac{\beta^2 \,\mathbb{E}[X_j^2 \,|\, X_\ell]}{1 + \beta^2 \,\mathbb{E}(X_j^2 \,|\, X_\ell)}}$$

Thus $\rho(X_k, X_j \,|\, X_\ell) \geq c \Leftrightarrow \beta^2 \,\mathbb{E}(X_j^2 \,|\, X_\ell) \geq \frac{c^2}{1-c^2}$. Since $X_{\phi_h} \perp\!\!\!\perp \eta_j \,|\, X_\ell$, we have $\mathbb{E}(X_j^2 \,|\, X_\ell) = 1 + \beta^2 \,\mathbb{E}(X_{\phi_h}^2 \,|\, X_\ell) \geq 1$, then $\beta^2 \,\mathbb{E}(X_j^2 \,|\, X_\ell) \geq \beta^2 = 2c^2 \geq \frac{c^2}{1-c^2}$ when $c^2 \leq 1/2$.

- For the third case, denote $v = \mathbb{E}[X_j^2]$, let's compute the covariance matrix of (X_k, X_k, X_ℓ):

$$\begin{pmatrix} v & \beta v & \beta^{h+2} v \\ \beta v & \beta^2 v + 1 & \beta^{h+1}(\beta^2 v + 1) \\ \beta^{h+2} v & \beta^{h+1}(\beta^2 v + 1) & \beta^{2(h+2)} v + \beta^{2(h+1)} + \cdots + \beta^2 + 1 \end{pmatrix}$$

Denote $V(v,h) = \beta^{2(h+2)} v + \beta^{2(h+1)} + \cdots + \beta^2 + 1$. The covariance matrix of (X_j, X_k) given X_ℓ is

$$\begin{aligned}
&\begin{pmatrix} v & \beta v \\ \beta v & \beta^2 v + 1 \end{pmatrix} - \frac{1}{V(v,h)} \begin{pmatrix} \beta^{2(h+2)} v^2 & \beta^{2h+3} v(\beta^2 v + 1) \\ \beta^{2h+3} v(\beta^2 v + 1) & \beta^{2(h+1)} (\beta^2 v + 1)^2 \end{pmatrix} \\
=&\frac{1}{V(v,h)} \Bigg[\begin{pmatrix} \beta^{2(h+2)} v^2 + \beta^{2(h+1)} v + \cdots + \beta^2 v + v & \beta^{2(h+2)+1} v^2 + \beta^{2(h+1)+1} v + \cdots + \beta^3 v + \beta v \\ \beta^{2(h+2)+1} v^2 + \beta^{2(h+1)+1} v + \cdots + \beta^3 v + \beta v & (\beta^{2(h+2)+2} v^2 + \beta^{2(h+1)+2} v + \cdots + \beta^4 v + \beta^2 v \\ & \quad + \beta^{2(h+2)} v + \beta^{2(h+1)} + \cdots + \beta^2 + 1) \end{pmatrix} \\
&- \begin{pmatrix} \beta^{2(h+2)} v^2 & \beta^{2h+3} v(\beta^2 v + 1) \\ \beta^{2h+3} v(\beta^2 v + 1) & \beta^{2(h+1)} (\beta^2 v + 1)^2 \end{pmatrix} \Bigg] \\
=&\frac{1}{V(v,h)} \begin{pmatrix} (\beta^{2(h+1)} + \beta^{2h} + \cdots + \beta^2 + 1) v & (\beta^{2h} + \beta^{2(h-1)} \cdots + \beta^2 + 1) \beta v \\ (\beta^{2h} + \beta^{2(h-1)} \cdots + \beta^2 + 1) \beta v & (\beta^{2h} + \beta^{2(h-1)} + \cdots + \beta^2 + 1)(\beta^2 v + 1) \end{pmatrix}
\end{aligned}$$

Thus the conditional correlation is

$$\begin{aligned}
\rho(X_j, X_k \,|\, X_\ell) &= \frac{\beta v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}{\sqrt{v \times \frac{1-\beta^{2(h+2)}}{1-\beta^2} \times (1 + \beta^2 v) \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}} \\
&= \sqrt{\frac{\beta^2 v}{1 + \beta^2 v} \times \frac{1 - \beta^{2(h+1)}}{1 - \beta^{2(h+2)}}}
\end{aligned}$$

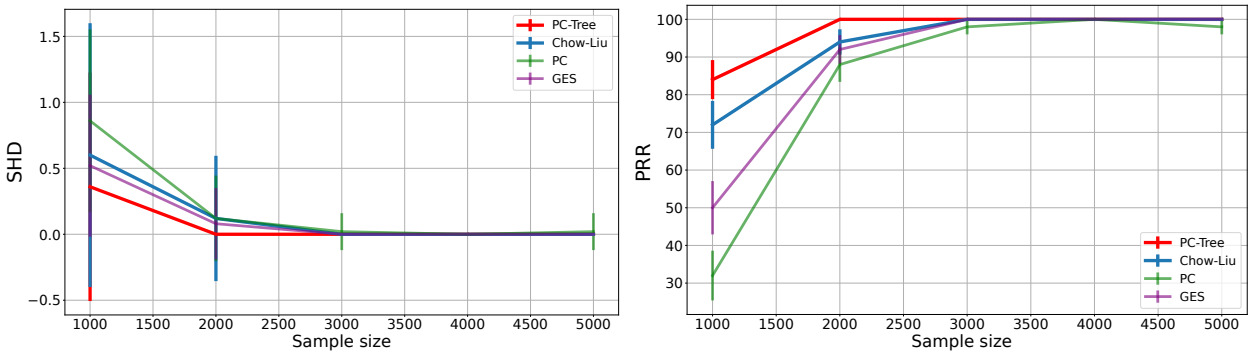
Denote $f(h) = \frac{1-\beta^{2(h+1)}}{1-\beta^{2(h+2)}} = 1 - \frac{(1-\beta^2)\beta^{2(h+1)}}{1-\beta^{2(h+2)}}$, which is increasing in h with minimum value being $f(0) = \frac{1}{1+\beta^2}$. Therefore, $\rho(X_j, X_k \,|\, X_\ell) \geq c \Leftrightarrow \beta^2 v \geq \frac{c^2}{f(h) - c^2}$. Since $v = \mathbb{E}[X_j^2] \geq 1$ for all $j \in [d]$, then $\beta^2 v \geq \beta^2 = 2c^2 \geq \frac{c^2}{1/(1+2c^2) - c^2} \geq \frac{c^2}{f(h) - c^2}$ when $c^2 \leq 1/5$, which yields the bound.

- For the forth case, analogously, denote $v = \mathbb{E}[X_j^2]$, let's compute the covariance matrix of (X_k, X_k, X_ℓ):

$$\begin{pmatrix} v & \beta v & \beta^{h+1} v \\ \beta v & \beta^2 v + 1 & \beta^{h+2} v \\ \beta^{h+1} v & \beta^{h+2} v & \beta^{2(h+1)} + \beta^{2h} + \cdots + \beta^2 + 1 \end{pmatrix}$$

Denote $W(v,h) = \beta^{2(h+1)} + \beta^{2h} v + + \cdots + \beta^2 + 1$. The covariance matrix of (X_j, X_k) given X_ℓ is

$$\begin{aligned}
&\begin{pmatrix} v & \beta v \\ \beta v & \beta^2 v + 1 \end{pmatrix} - \frac{1}{W(v,h)} \begin{pmatrix} \beta^{2(h+1)} v^2 & \beta^{2h+3} v^2 \\ \beta^{2h+3} v^2 & \beta^{2(h+2)} v^2 \end{pmatrix} \\
=&\frac{1}{W(v,h)} \Bigg[\begin{pmatrix} \beta^{2(h+1)} v^2 + \beta^{2h} v + \cdots + \beta^2 v + v & \beta^{2(h+1)+1} v^2 + \beta^{2h+1} v + \cdots + \beta^3 v + \beta v \\ \beta^{2(h+1)+1} v^2 + \beta^{2h+1} v + \cdots + \beta^3 v + \beta v & (\beta^{2(h+1)+2} v^2 + \beta^{2h+2} v + \cdots + \beta^4 v + \beta^2 v \\ & \quad + \beta^{2(h+1)} v + \beta^{2h} + \cdots + \beta^2 + 1) \end{pmatrix} \\
&- \begin{pmatrix} \beta^{2(h+1)} v^2 & \beta^{2h+3} v^2 \\ \beta^{2h+3} v^2 & \beta^{2(h+2)} v^2 \end{pmatrix} \Bigg] \\
=&\frac{1}{W(v,h)} \begin{pmatrix} (\beta^{2h} + \beta^{2(h-1)} + \cdots + \beta^2 + 1) v & (\beta^{2h} + \beta^{2(h-1)} \cdots + \beta^2 + 1) \beta v \\ (\beta^{2h} + \beta^{2(h-1)} \cdots + \beta^2 + 1) \beta v & (\beta^{2h} + \beta^{2(h-1)} + \cdots + \beta^2 + 1)(\beta^2 v + 1) + \beta^{2h+2} v \end{pmatrix}
\end{aligned}$$



(a) SHD (b) PRR

Figure 6: SHD and PRR for Gaussian η and $d = 10$.

Thus the conditional correlation is

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{\beta v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}{\sqrt{v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2} \times \left[(1+\beta^2 v) \times \frac{1-\beta^{2(h+1)}}{1-\beta^2} + \beta^{2h} \times \beta^2 v\right]}} = \sqrt{\frac{\beta^2 v}{1 + \left(1 + \frac{\beta^{2h}}{g(h)}\right)\beta^2 v}}$$

where $g(h) = \frac{1-\beta^{2(h+1)}}{1-\beta^2} \geq 1$ for $h \geq 0$. Since $\beta^2 = 2c^2 \leq 1$, then $1 + \beta^{2h}/g(h) \leq 2$. Since $\rho(X_j, X_k \,|\, X_\ell) \geq c \Leftrightarrow \beta^2 v \geq \frac{c^2}{1-\left(1+\frac{\beta^{2h}}{g(h)}\right)c^2}$, and $v = \mathbb{E}[X_j^2] \geq 1$ for all $j \in [d]$, then $\beta^2 v \geq \beta^2 = 2c^2 \geq \frac{c^2}{1-2c^2} \geq \frac{c^2}{1-\left(1+\frac{\beta^{2h}}{g(h)}\right)c^2}$ when $c^2 \leq 1/5$, which completes the proof.

□

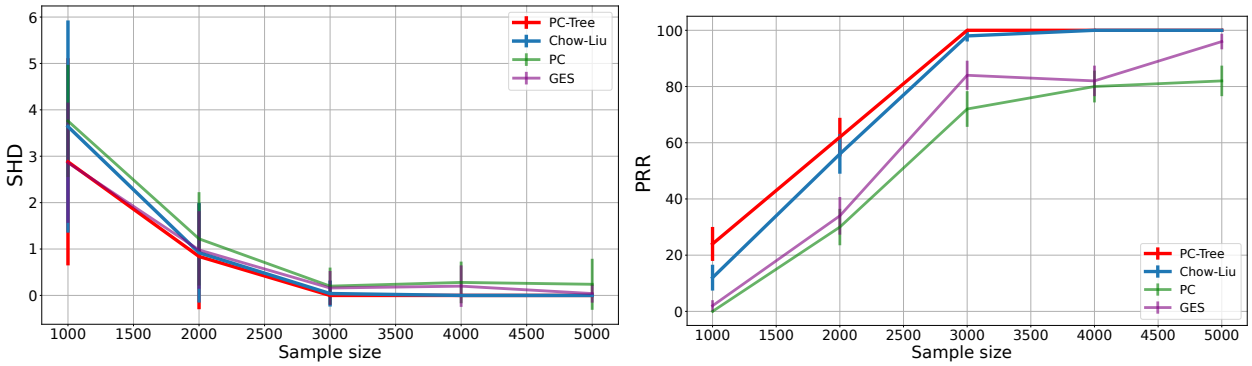
D EXPERIMENTS

Synthetic Data Generation We generate trees using package `networkx`, then randomly pick a node as root and orient it into a directed tree. We consider number of nodes $d \in \{10, 50, 100\}$. To generate the data as in (2.2), we uniformly sample β_k from the interval $(-0.5, 0.1] \cup [0.1, 0.5)$ as our coefficient weight. For sample size $n = \{1000, 2000, 3000, 4000, 5000\}$, we generate our i.i.d. samples $X \in \mathbb{R}^{n \times d}$ according to (2.2), where $\eta \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$. Besides, we also present experiments on agnostic setting where $\eta \sim \mathcal{U}(-1, 1)$ is uniform distribution, or $\eta \sim$ Laplace$(0, 1)$ is Laplace distribution.

Baselines We have employed two baseline algorithms: the PC algorithm has been executed using the Python package `Causal-learn`, while the GES algorithm has been implemented with `py-tetrad`.

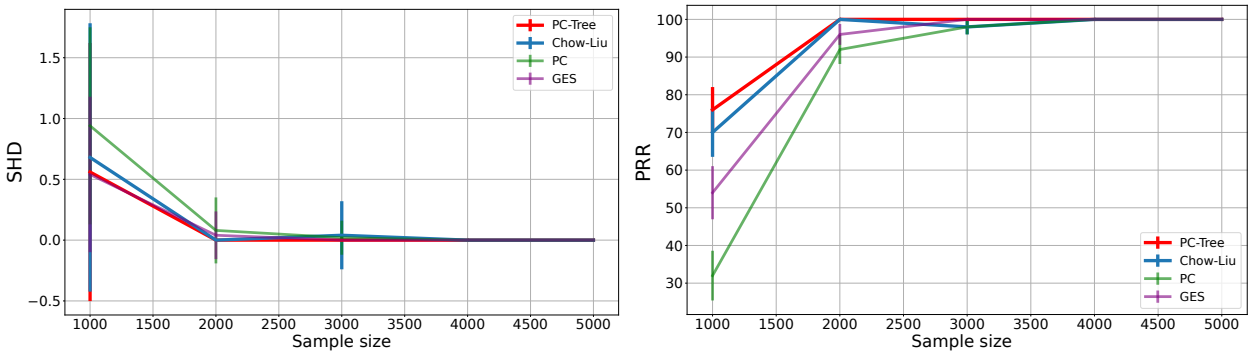
Evaluation For each experiment setup, we report the average (over 50 random instantiations) Structural Hamming Distance (SHD) between the ground truth and our estimated graph skeleton, and the Precise Recovery Rate (PRR), which is the frequency of exact recovery of the tree skeleton. Results are reported in Figure 6-13. All experiments were conduced on an Intel Core i7-12800H 2.40GHz CPU.

Agnostic Learning Additionally, we investigated the algorithm's performance under conditions where the assumption is violated. Specifically, we examined the impact on our algorithm's performance when the coefficients β_k in (2.2) are not independently and identically distributed (i.i.d.). To address this question, we conducted agnostic learning experiments and present the corresponding results.



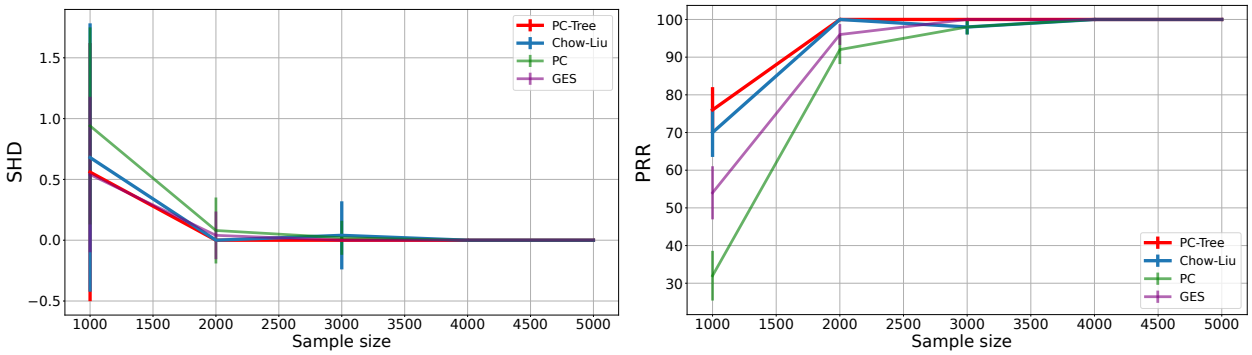
(a) SHD (b) PRR

Figure 7: SHD and PRR for Gaussian η and $d = 50$.



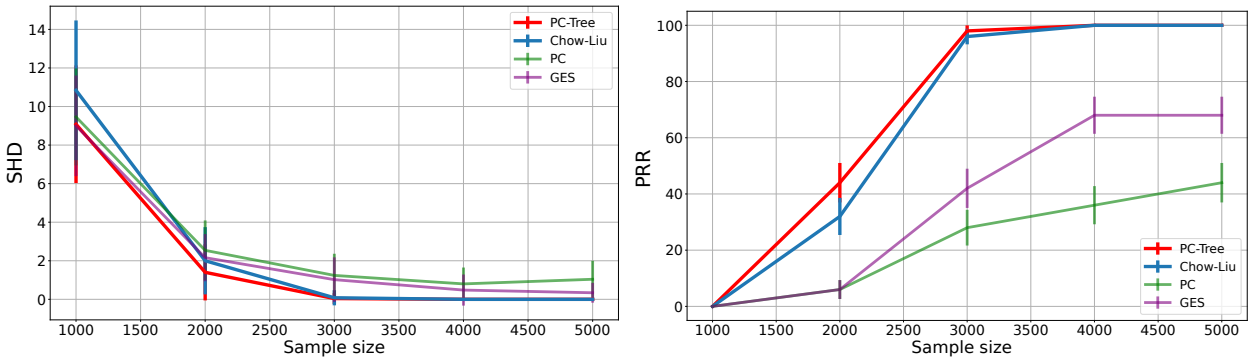
(a) SHD (b) PRR

Figure 8: SHD and PRR for Uniform η and $d = 10$.



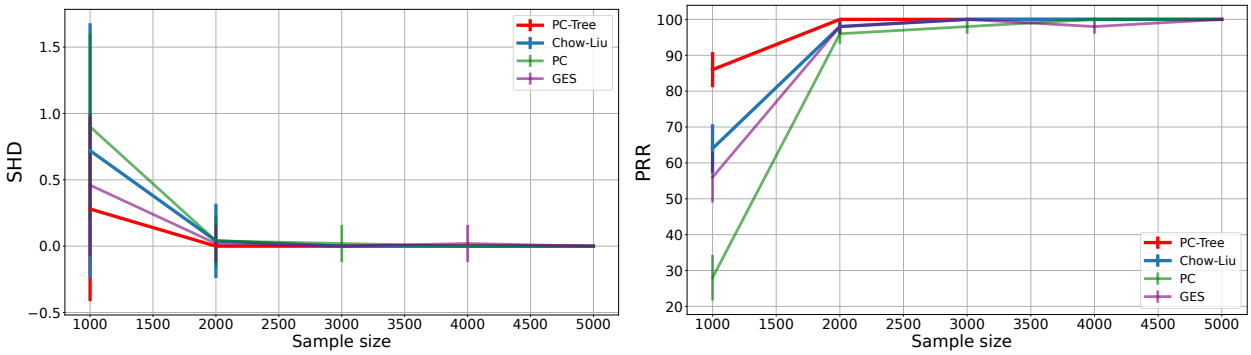
(a) SHD (b) PRR

Figure 9: SHD and PRR for Uniform η and $d = 50$.



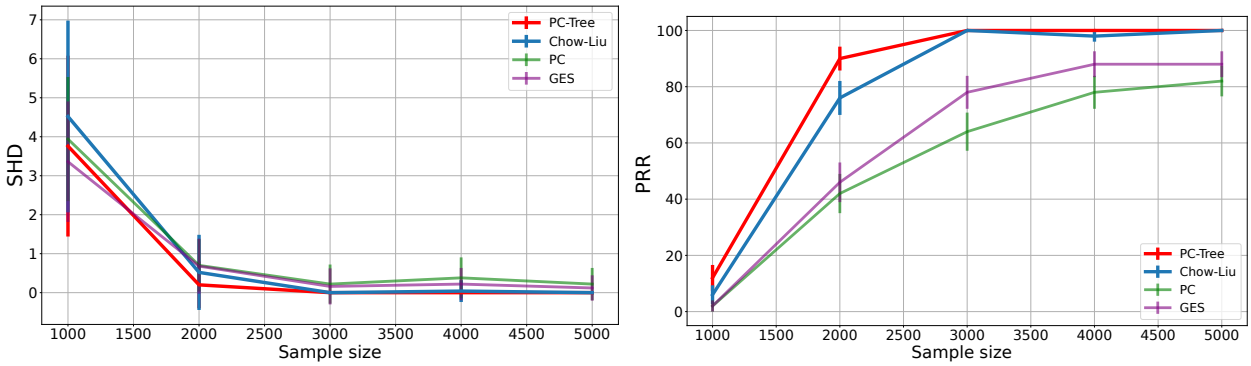
(a) SHD (b) PRR

Figure 10: SHD and PRR for Uniform η and $d = 100$.



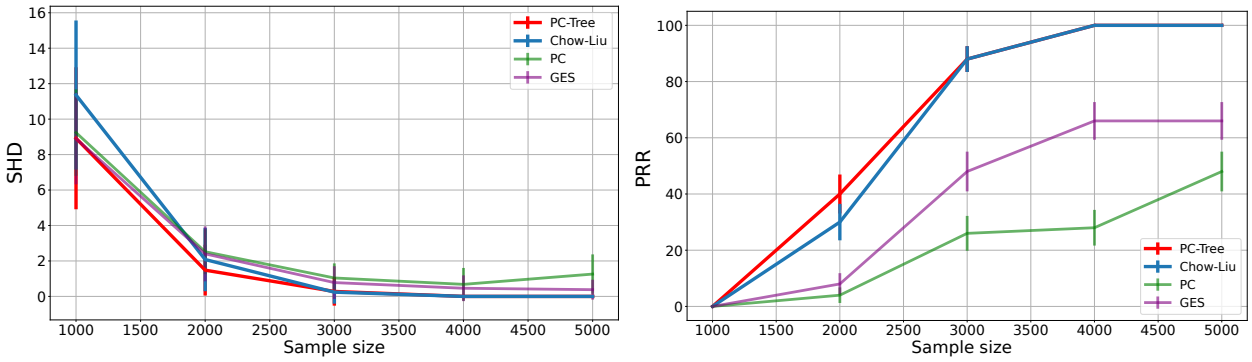
(a) SHD (b) PRR

Figure 11: SHD and PRR for Laplace η and $d = 10$.



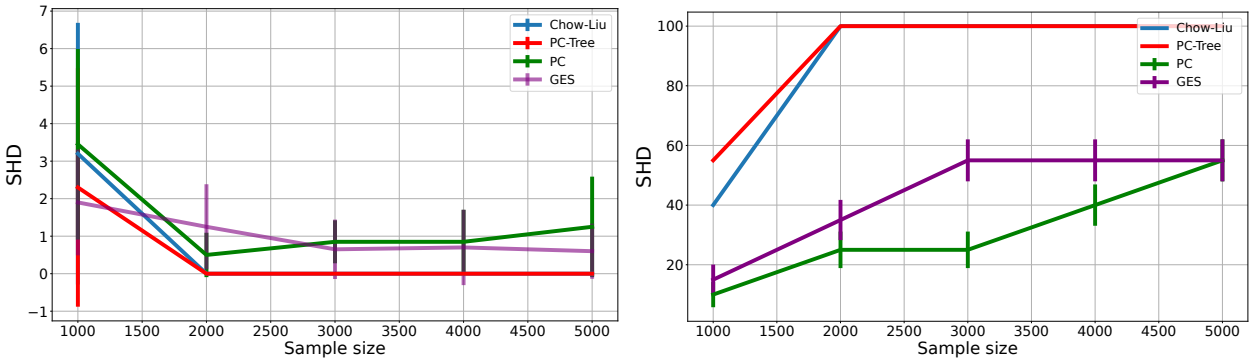
(a) SHD (b) PRR

Figure 12: SHD and PRR for Laplace η and $d = 50$.



(a) SHD (b) PRR

Figure 13: SHD and PRR for Laplace η and $d = 100$.



(a) SHD comparison (non-i.i.d.) (b) PRR comparison (non-i.i.d.)

Figure 14: Performance comparison for PC-Tree, Chow-Liu, PC and GES algorithm evaluated on SHD and PRR in (a) and (b) for non-iid β_k. The red, blue, green, purple lines are for PC-Tree, Chow-Liu, PC and GES respectively.

See Figure 14 for results with non-iid β_k. Specifically, $\beta_k = \alpha_k + z$, where we sample α_k iid uniformly and z uniformly, applying the same z to all α_k. Here, z introduces dependence among β_k. When $z = 0$, β_k is i.i.d., and when $z \neq 0$, β_k is non-i.i.d. For brevity, we only report the most relevant setting with $d = 100$ nodes and data are Gaussian. We simulated random directed trees and synthetic data via equation (2.2). We can see the performance of both PC-tree and Chow-Liu are less affected even when β_k are non i.i.d: The Structural Hamming Distance (SHD) becomes 0 in both i.i.d and non i.i.d. setting, and the Precise Recovery Rate (PRR) also outperforms other methods.