Journal of Machine Learning Research 24 (2023) 1-48 Submitted 7/21; Revised 1/23; Published 2/23

Inference for a Large Directed Acyclic Graph with Unspecified Interventions

Chunlin Li* LI000007@UMN.EDU
School of Statistics,
University of Minnesota,
Minneapolis, MN 55455, USA

Xiaotong Shen XSHEN@UMN.EDU
School of Statistics,
University of Minnesota,
Minneapolis, MN 55455, USA

Wei Pan PANXX014@UMN.EDU
Division of Biostatistics,
University of Minnesota,
Minneapolis, MN 55455, USA

Editor: Pradeep Ravikumar

Abstract

Statistical inference of directed relations given some unspecified interventions (i.e., the intervention targets are unknown) is challenging. In this article, we test hypothesized directed relations with unspecified interventions. First, we derive conditions to yield an identifiable model. Unlike classical inference, testing directed relations requires identifying the ancestors and relevant interventions of hypothesis-specific primary variables. To this end, we propose a peeling algorithm based on nodewise regressions to establish a topological order of primary variables. Moreover, we prove that the peeling algorithm yields a consistent estimator in low-order polynomial time. Second, we propose a likelihood ratio test integrated with a data perturbation scheme to account for the uncertainty of identifying ancestors and interventions. Also, we show that the distribution of a data perturbation test statistic converges to the target distribution. Numerical examples demonstrate the utility and effectiveness of the proposed methods, including an application to infer gene regulatory networks. The R implementation is available at `https://github.com/chunlinli/intdag`.

Keywords: high-dimensional inference, data perturbation, structure learning, peeling algorithm, identifiability

1. Introduction

Directed relations are essential to explaining pairwise dependencies among multiple interacting units. In gene network analysis, regulatory gene-to-gene relations are a focus of biological investigation (Sachs et al., 2005), while in a human brain network, scientists investigate causal influences among regions of interest to understand how the brain functions (Liu et al., 2017). In such a situation, a Gaussian directed acyclic graph (DAG)

*. To whom correspondence should be addressed.

is commonly employed to describe the directed relations; however, inferring the directed effects without other information is generally impossible because a Gaussian DAG often lacks model identifiability (van de Geer and Bühlmann, 2013). Hence, external interventions are introduced to treat a non-identifiable situation (Heinze-Deml et al., 2018). For instance, the genetic variants such as single-nucleotide polymorphisms (SNPs) can be, and indeed are increasingly, treated as external interventions to infer inter-trait causal relations in a quantitative trait network (Brown and Knowles, 2020) and gene interactions in a gene regulatory network (Teumer, 2018; Molstad et al., 2021). In neuroimaging analysis, scientists use randomized experimental stimuli as interventions to identify causal relations in a functional brain network (Grosse-Wentrup et al., 2016; Bergmann and Hartwigsen, 2021). However, the interventions in these studies often have unknown targets and off-target effects (Jackson et al., 2003; Eaton and Murphy, 2007). Consequently, inferring directed relations while identifying useful interventions for inference is critical. This paper focuses on the simultaneous inference of directed relations subject to unspecified interventions (i.e., the intervention targets are unknown).

In a DAG model, the research has been centered on the reconstruction of directed relations in observational and interventional studies (van de Geer and Bühlmann, 2013; Oates et al., 2016; Zheng et al., 2018; Yuan et al., 2019; Li et al., 2023); see Heinze-Deml et al. (2018) for a review. For uncertainty quantification, Bayesian methods (Friedman and Koller, 2003; Luo and Zhao, 2011; Viinikka et al., 2020) have been popular. Yet, statistical inference remains under-studied, especially for interventional models in high dimensions (Peters et al., 2016; Rothenhäusler et al., 2019). Recently, for observational data, Janková and van de Geer (2018) propose a debiased test of a single directed relation, and Li et al. (2020) derive a constrained likelihood ratio test for multiple directed relations.

Despite progress, challenges remain. First, inferring directed relations requires identifying a certain DAG topological order (van de Geer and Bühlmann, 2013), while the identifiability in a Gaussian DAG with unspecified interventions remains under-explored. Second, the inferential results should agree with the acyclicity requirement for a DAG. As a result, degenerate and intractable situations can occur, making inference greatly different from the classical ones. Third, likelihood-based methods for learning the DAG topological order often use permutation search (van de Geer and Bühlmann, 2013) or continuous optimization subject to the acyclicity constraint (Zheng et al., 2018; Yuan et al., 2019), where a theoretical guarantee of the actual estimate (instead of the global optimum) has not been established for these approaches. Recently, an important line of work (Ghoshal and Honorio, 2018; Rajendran et al., 2021; Rolland et al., 2022) has focused on order-based algorithms with computational and statistical guarantees. However, in Gaussian DAGs, existing methods often rely on some error scale assumptions (Peters and Bühlmann, 2014), which is sensitive to variable scaling like the common practice of standardizing variables. This drawback could limit their applications, especially in causal inference, as causal relations are typically invariant to scaling.

To address the above issues, we develop structure learning and inference methods for a Gaussian DAG with unspecified additive interventions. Unlike the existing approach treating structure discovery and subsequent inference separately, our proposal integrates DAG structure learning and testing of directed relations, accounting for the uncertainty of structure learning for inference. With suitable interventions called instrumental variables (IVs),

the proposed approach removes the restrictive error scale assumptions and delivers creditable outcomes with theoretical guarantees in low-order polynomial time. This indicates IVs, a well-known tool in causal inference (Angrist et al., 1996), can play important roles in structure learning even if some interventions do not meet the IV criteria. Our contributions are summarized as follows.

- For modeling, we establish the identifiability conditions for a Gaussian DAG with unspecified interventions. In particular, the conditions allow interventions on more than one target, which is suitable for multivariate causal analysis (Murray, 2006).
- For methodology, we develop likelihood ratio tests for directed edges and pathways in a super-graph of the true DAG, called the ancestral relation graph (ARG), where the ARG is formed by ancestral relations and candidate interventional relations, offering the topological order for inference. We reconstruct the ARG by the peeling algorithm, which automatically meets the acyclicity requirement. On this basis, we introduce the concepts of nondegeneracy and regularity to characterize the behavior of hypothesis testing under a DAG model. By integrating structure learning with inference, we account for the uncertainty of ARG estimation for the proposed tests via a novel data perturbation (DP) scheme, which effectively controls the type-I error while enjoying high statistical power.
- For theory, we prove that the proposed peeling algorithm based on nodewise regressions yields consistent results in $O(p \times \log \kappa^{\circ}_{\max} \times (q^3 + nq^2))$ operations almost surely, where p, q are the numbers of primary and intervention variables, n is the sample size, and $\kappa^{\circ}_{\max}$ is the sparsity. Then we justify the proposed DP inference method by establishing the convergence of the DP likelihood ratio to the target distribution and desired power properties.
- The numerical studies and real data analysis demonstrate the utility and effectiveness of the proposed methods. The implementation of the proposed tests and structure learning method is available at `https://github.com/chunlinli/intdag`.

The rest of the article is structured as follows. Section 2 establishes model identifiability and states two inference problems of interest. Section 3 develops the proposed methods for structure learning and statistical inference. Section 4 presents statistical theory to justify the proposed methods. Section 5 performs simulation studies, followed by an application to infer gene pathways from gene expression and SNP data in Section 6. Section 7 concludes the article. The Appendix contains illustrative examples and technical proofs.

2. Gaussian Directed Acyclic Graph with Additive Interventions

To infer directed relations among p primary variables (i.e., variables of primary interest) $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$, consider a structural equation model with q additive interventions:

$$\boldsymbol{Y} = \mathbf{U}^\top \boldsymbol{Y} + \mathbf{W}^\top \boldsymbol{X} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2), \tag{1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$ is a vector of additive intervention variables, $\mathbf{U} \in \mathbb{R}^{p \times p}$ and $\mathbf{W} \in \mathbb{R}^{q \times p}$ are unknown coefficient matrices, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^\top$ is a vector of random

errors with $\sigma_j^2 > 0$; $j = 1, \ldots, p$. In (1), $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$ but the components of $\boldsymbol{X}$ can be dependent. The matrix $\mathbf{U}$ specifies the directed relations among $\boldsymbol{Y}$, where $\mathrm{U}_{kj} \neq 0$ if Y_k is a direct cause of Y_j, denoted by $Y_k \to Y_j$, and Y_k is called a parent of Y_j or Y_j a child of Y_k. Thus, $\mathbf{U}$ represents a directed graph, which is further required to be acyclic to ensure the validity of the local Markov property (Spirtes et al., 2000). The matrix $\mathbf{W}$ specifies the targets and strengths of interventions, where $\mathrm{W}_{lj} \neq 0$ indicates X_l intervenes on Y_j, denoted by $X_l \to Y_j$. In (1), no directed edge from a primary variable Y_j to an intervention variable X_l is permissible.

In what follows, we will focus on the DAG $\mathcal{G} = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}, \mathcal{I})$ with primary variables $\boldsymbol{Y}$, intervention variables $\boldsymbol{X}$, primary edges $\mathcal{E} = \{(k, j) : \mathrm{U}_{kj} \neq 0\}$, and intervention edges $\mathcal{I} = \{(l, j) : \mathrm{W}_{lj} \neq 0\}$. To facilitate discussion, we introduce some concepts and notations for $\mathcal{G}$. If there is a directed path $Y_k \to \cdots \to Y_j$ in $\mathcal{G}$, Y_k is an ancestor of Y_j or Y_j is a descendant of Y_k. If $Y_k \to Y_j$ and there is no other directed path from Y_k to Y_j, then we say Y_k is an unmediated parent of Y_j. Let $\textsc{pa}_{\mathcal{G}}(j) = \{k : Y_k \to Y_j\}$, $\textsc{an}_{\mathcal{G}}(j) = \{k : Y_k \to \cdots \to Y_j\}$, and $\textsc{in}_{\mathcal{G}}(j) = \{l : X_l \to Y_j\}$ be the parent, ancestor, and intervention sets of Y_j, respectively.

2.1 Identifiability and Instruments

Model (1) is generally non-identifiable without interventions ($\mathbf{W} = \mathbf{0}$) when errors do not meet some requirements such as the equal-variance assumption (Peters and Bühlmann, 2014) and its variants (Ghoshal and Honorio, 2018; Rajendran et al., 2021). Moreover, the model can be identified when $\boldsymbol{\varepsilon}$ in (1) is replaced by non-Gaussian errors (Shimizu et al., 2006) or linear relations are replaced by nonlinear ones (Peters et al., 2014). Regardless, suitable interventions can make (1) identifiable. When intervention targets are known, the identifiability issue has been studied (Oates et al., 2016; Chen et al., 2018). However, it is less so when the exact targets and strengths of interventions are unknown as in many biological applications (Jackson et al., 2003; Kulkarni et al., 2006), which is referred to as the case of *unspecified* or *uncertain* interventions (Heinze-Deml et al., 2018; Eaton and Murphy, 2007; Squires et al., 2020).

We now categorize interventions as instruments and invalid instruments.

Definition 1 (DAG instrument) *An intervention variable is an instrument in $\mathcal{G}$ if*

(A) it intervenes on at least one primary variable in $\mathcal{G}$;

(B) it does not intervene on more than one primary variable in $\mathcal{G}$.

Otherwise, it is an invalid instrument in $\mathcal{G}$.

Here, (A) requires an intervention to be active, while (B) prevents simultaneous interventions of a single intervention variable on multiple primary variables. This is critical to identifiability because an instrument on a (potential) cause variable Y_1 helps reveal its directed effect on an outcome variable Y_2, which breaks the symmetry in a Gaussian DAG that results in non-identifiability of directed relations $Y_1 \to Y_2$ and $Y_2 \to Y_1$.

Remark 2 *The conventional definition of instrumental variable differs from Definition 1. In the literature (Angrist et al., 1996), an instrument X for estimating the effect from a*

potential cause Y_1 to the outcome Y_2 is required to satisfy (i) X is related to Y_1, called relevance, (ii) X has no directed edge to the outcome Y_2, called exclusion, and (iii) X is not related to unmeasured confounders, called unconfoundedness. In Definition 1, (A) is the relevance property, (B) generalizes the exclusion property for a DAG model, and the unconfoundedness is satisfied because no confounder is present in model (1).

Next, we make some assumptions on intervention variables to yield an identifiable model, where dependencies among intervention variables are permissible.

Assumption 1 *Assume that model* (1) *satisfies the following conditions.*

(1A) $\mathbb{E}\,\boldsymbol{X}\boldsymbol{X}^\top$ *is positive definite.*

(1B) $\mathrm{Cov}(Y_j, X_l \mid \boldsymbol{X}_{\{1,\ldots,q\}\setminus\{l\}}) \neq 0$ *if* X_l *intervenes on any unmediated parent of* Y_j.

(1C) Each primary variable is intervened by at least one instrument.

Assumption 1A imposes mild distributional restrictions on $\boldsymbol{X}$, permitting discrete variables such as SNPs. Assumption 1B requires the interventional effects through unmediated parents not to cancel out, as multiple targets from an invalid instrument are permitted. Importantly, if either Assumption 1B or 1C fails, model (1) is generally not identifiable, as shown in Example 2 of Appendix A.1. In Section 5, we empirically examine the situation when Assumption 1C is not met.

Proposition 3 *Under Assumption 1,* $(\mathbf{U}, \mathbf{W}, \mathbf{\Sigma})$ *in model* (1) *are identifiable from the distribution of* $(\boldsymbol{Y}, \boldsymbol{X})$.

Proposition 3 (proved in Appendix B.1) is derived for a DAG model with unspecified interventions. This is in contrast to Proposition 1 of Chen et al. (2018), which proves the identifiability of the parameters in a directed graph with target-known instruments on each primary variable. Moreover, the estimated graph in Chen et al. (2018) may be cyclic and lacks the local Markov property for causal interpretation (Spirtes et al., 2000).

2.2 Problem Statement: Inference for a DAG

Our goal is to perform statistical inference of directed edges and pathways in the DAG $\mathcal{G}$. Let $\mathcal{H} \subseteq \{(k,j) : k \neq j, 1 \leq k, j \leq p\}$ be an edge set among primary variables $\{Y_1, \ldots, Y_p\}$, where $(k,j) \in \mathcal{H}$ specifies a (hypothesized) directed edge $Y_k \to Y_j$ in (1). We shall focus on two types of testing with null and alternative hypotheses H_0 and H_a. For simultaneous testing of directed edges,

$$H_0 : \mathrm{U}_{kj} = 0; \text{ for all } (k,j) \in \mathcal{H} \quad \text{versus} \quad H_a : \mathrm{U}_{kj} \neq 0 \text{ for some } (k,j) \in \mathcal{H}; \tag{2}$$

for simultaneous testing of directed pathways,

$$H_0 : \mathrm{U}_{kj} = 0; \text{ for some } (k,j) \in \mathcal{H} \quad \text{versus} \quad H_a : \mathrm{U}_{kj} \neq 0 \text{ for all } (k,j) \in \mathcal{H}, \tag{3}$$

where $(\mathbf{U}_{\mathcal{H}^c}, \mathbf{W}, \mathbf{\Sigma})$ are unspecified nuisance parameters and ^{c} is the set complement. Note that H_0 in (3) is a composite hypothesis that can be decomposed into sub-hypotheses

$$H_{0,\nu} : \mathrm{U}_{k_\nu, j_\nu} = 0, \quad \text{versus} \quad H_{a,\nu} : \mathrm{U}_{k_\nu, j_\nu} \neq 0; \quad \nu = 1, \ldots, |\mathcal{H}|,$$

where $\mathcal{H} = \{(k_1, j_1), \ldots, (k_{|\mathcal{H}|}, j_{|\mathcal{H}|})\}$ and testing each sub-hypothesis is a directed edge test. Thus, we treat (3) as an extension of (2).

We will also estimate $(\mathbf{U}, \mathbf{W})$ as well as identify the nonzero elements in $\mathbf{U}$ to recover the directed edges among the primary variables $\boldsymbol{Y}$ in $\mathcal{G}$.

3. Methodology

This section develops the main methodology, including the peeling algorithm for structure learning and the data perturbation inference for simultaneous testing of directed edges (2) and pathways (3). To proceed, suppose the data matrices $\mathbf{Y}_{n\times p} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^\top$ and $\mathbf{X}_{n\times q} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ are given, where the rows $\{(\boldsymbol{Y}_i^\top, \boldsymbol{X}_i^\top)\}_{1\le i\le n}$ are independently sampled from (1). Then the log-likelihood is (up to a constant)

$$L(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = -\frac{1}{2}\sum_{i=1}^{n} \left\|\boldsymbol{\Sigma}^{-1/2}((\mathbf{I} - \mathbf{U}^\top)\boldsymbol{Y}_i - \mathbf{W}^\top \boldsymbol{X}_i)\right\|_2^2 - n\log\sqrt{\det(\boldsymbol{\Sigma})}, \tag{4}$$

where $\boldsymbol{\theta} = (\mathbf{U}, \mathbf{W})$, $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$, and $\mathbf{U}$ is subject to the acyclicity constraint (Zheng et al., 2018; Yuan et al., 2019) in that no directed cycle is permissible in the DAG.

One major challenge to this likelihood approach lies in the optimization of (4) subject to the acyclicity constraint, which imposes difficulty on not only computation but also asymptotic theory. As a result, there is a gap between the asymptotic distribution of a global maximum and that of the actual estimate which can be a local maximum (Janková and van de Geer, 2018; Li et al., 2020). Moreover, the actual estimate may give an imprecise topological order, tending to impact adversely on inference.

To circumvent the acyclicity requirement, we propose to use the ancestral relation graph (ARG) to describe the topological order of the DAG, where the ARG can be efficiently estimated without explicitly imposing the acyclicity constraint while enjoying a statistical guarantee of the actual estimate.

Definition 4 (Ancestral relation graph)

(A) A graph $\mathcal{M} = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{A}, \mathcal{C})$ is an ARG if it is acyclic and $\mathcal{A} = \{(k, j) : k \in \mathrm{AN}_{\mathcal{M}}(j)\}$.

(B) Given DAG $\mathcal{G} = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}, \mathcal{I})$, its ARG is defined as $\mathcal{G}_+ = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}_+, \mathcal{I}_+)$, where

$$\mathcal{E}_+ = \Big\{(k, j) : k \in \mathrm{AN}_{\mathcal{G}}(j)\Big\}, \quad \mathcal{I}_+ = \Big\{(l, j) : l \in \bigcup_{k\in \mathrm{AN}_{\mathcal{G}}(j)\cup\{j\}} \mathrm{IN}_{\mathcal{G}}(k)\Big\}.$$

Given $\mathcal{G}_+$ (which is acyclic), we have $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{G}_+}, \boldsymbol{\theta}_{\mathcal{G}_+^c})$, where $\boldsymbol{\theta}_{\mathcal{G}_+} = (\mathbf{U}_{\mathcal{E}_+}, \mathbf{W}_{\mathcal{I}_+})$ are the (effective) parameters and $\boldsymbol{\theta}_{\mathcal{G}_+^c} = (\mathbf{U}_{\mathcal{E}_+^c}, \mathbf{W}_{\mathcal{I}_+^c}) = \mathbf{0}$. Then the log-likelihood (4) becomes

$$L((\boldsymbol{\theta}_{\mathcal{G}_+}, \mathbf{0}), \boldsymbol{\Sigma}) = -\sum_{j=1}^{p} \underbrace{\sum_{i=1}^{n} \Big(Y_{ij} - \sum_{(k,j)\in\mathcal{E}_+} U_{kj}Y_{ik} - \sum_{(l,j)\in\mathcal{I}_+} W_{lj}X_{il}\Big)^2}_{:=\mathrm{RSS}_j(\boldsymbol{\theta})} /2\sigma_j^2 + n\log(\sigma_j), \tag{5}$$

which involves $|\mathcal{E}_+| + |\mathcal{I}_+|$ parameters for $\boldsymbol{\theta}_{\mathcal{G}_+}$. From (5), we can reconstruct $\mathcal{G}$ and conduct inference for (2) and (3).

Our plan is as follows. In Section 3.1, we construct $\mathcal{G}_+$ without the acyclicity constraint for $\mathbf{U}$. On this basis, in Section 3.2 we develop likelihood ratio tests for (2) and (3).

3.1 Structure Learning via Peeling

This section develops a novel structure learning method to construct $\mathcal{G}_+$ in a hierarchical manner. First, we observe an important connection between primary variables and intervention variables. Rewrite (1) as

$$\boldsymbol{Y} = \mathbf{V}^\top \boldsymbol{X} + \boldsymbol{\varepsilon}_V, \quad \boldsymbol{\varepsilon}_V = (\mathbf{I} - \mathbf{U}^\top)^{-1}\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{-1}), \tag{6}$$

where $\boldsymbol{\Omega} = (\mathbf{I} - \mathbf{U})\boldsymbol{\Sigma}^{-1}(\mathbf{I} - \mathbf{U}^\top)$ and $\mathbf{V} = \mathbf{W}(\mathbf{I} - \mathbf{U})^{-1}$.

Proposition 5 *Suppose Assumption 1 is satisfied.*

(A) If $\mathrm{V}_{lj} \neq 0$*, then* X_l *intervenes on* Y_j *or an ancestor of* Y_j*;*

(B) In $\mathcal{G}$*,* Y_j *is a leaf variable (having no child) if and only if there is an instrument* X_l *such that* $\mathrm{V}_{lj} \neq 0$ *and* $\mathrm{V}_{lj'} = 0$ *for* $j' \neq j$*.*

The proof of Proposition 5 is deferred to Appendix B.2. Intuitively, $\mathrm{V}_{lj} \neq 0$ implies the dependence of Y_j on X_l through a directed path $X_l \to \cdots \to Y_j$, and hence that X_l intervenes on Y_j or an ancestor of Y_j. Thus, the instruments on a leaf variable are independent of the other primary variables conditional on the rest of interventions. This observation suggests a method to reconstruct the DAG topological order by recursively identifying and removing (i.e., peeling) the leaf variables.

Next, we discuss the estimation of $\mathbf{V}$ and construction of $\mathcal{G}_+$.

3.1.1 Nodewise constrained regressions

We estimate $\mathbf{V} = (\mathbf{V}_{\cdot 1}, \ldots, \mathbf{V}_{\cdot p})$ via nodewise ℓ_0-constrained regressions,

$$\widehat{\mathbf{V}}_{\cdot j} = \underset{\mathbf{V}_{\cdot j}}{\arg\min} \ \sum_{i=1}^{n} \left(Y_{ij} - \mathbf{V}_{\cdot j}^\top \boldsymbol{X}_i\right)^2 \quad \text{s.t.} \quad \sum_{l=1}^{q} \mathrm{I}(\mathrm{V}_{lj} \neq 0) \leq \kappa_j; \quad j = 1, \ldots, p, \tag{7}$$

where $1 \leq \kappa_j \leq q$ is an integer-valued tuning parameter controlling the sparsity and can be chosen by BIC or cross-validation. To solve (7), we use $J(z; \tau_j) = \min(|z|/\tau_j, 1)$ as a surrogate of $\mathrm{I}(z \neq 0)$ (Shen et al., 2012) and develop a difference-of-convex (DC) program with the ℓ_0-projection to improve the globality of the solution of (7). Specifically, at the $(t+1)$th iteration, given $\widetilde{\mathbf{V}}_{\cdot j}^{[t]}$, we solve the weighted Lasso problem,

$$\widetilde{\mathbf{V}}_{\cdot j}^{[t+1]} = \underset{\mathbf{V}_{\cdot j}}{\arg\min} \ \sum_{i=1}^{n} \left(Y_{ij} - \mathbf{V}_{\cdot j}^\top \boldsymbol{X}_i\right)^2 + 2n\gamma_j\tau_j \sum_{l=1}^{q} \mathrm{I}\left(|\widetilde{\mathrm{V}}_{lj}^{[t]}| \leq \tau_j\right)|\mathrm{V}_{lj}|; \quad j = 1, \ldots, p, \tag{8}$$

where $\gamma_j > 0$ is an internal hyperparameter used by the DC program; see Remark 6 below. The DC program terminates at $\widetilde{\mathbf{V}}_{\cdot j} = \widetilde{\mathbf{V}}_{\cdot j}^{[t]}$ such that $\|\widetilde{\mathbf{V}}_{\cdot j}^{[t+1]} - \widetilde{\mathbf{V}}_{\cdot j}^{[t]}\|_\infty \leq \sqrt{\mathrm{tol}}$ or t achieves the maximum iteration number, where tol is the machine precision. Then, the solution $\widehat{\mathbf{V}}_{\cdot j}$ of (7) is computed by projecting $\widetilde{\mathbf{V}}_{\cdot j}$ onto the set $\left\{\mathbf{v} \in \mathbb{R}^q : \|\mathbf{v}\|_0 \leq \kappa_j\right\}$.

Algorithm 1 summarizes the computation method.

Algorithm 1: Constrained estimation via DC program + ℓ_0 projection

Input: data $\mathbf{Y}$ and $\mathbf{X}$.

Parameters: $\{(\kappa_j, \tau_j)\}_{1\le j\le p}$; candidate values $\{\gamma^{(1)} > \cdots > \gamma^{(R)}\}$ for γ_j.

Output: the estimate $\widehat{\mathbf{V}}$.

for *each* $1 \le j \le p$ **do**
  for *each* $1 \le r \le R$ **do**
    Initialize $\widehat{\mathbf{V}}_{\cdot j}^{(r)} \leftarrow \mathbf{0}$, and $\gamma_j \leftarrow \gamma^{(r)}$;
    Compute $\widetilde{\mathbf{V}}_{\cdot j}$ via DC program (8) with $\widetilde{\mathbf{V}}_{\cdot j}^{[0]} \leftarrow \mathbf{0}$;
    Let $B \leftarrow \big\{l : |\widetilde{\mathrm{V}}_{lj}| \neq 0$ is among the largest κ_j elements of $\{|\widetilde{\mathrm{V}}_{l'j}|\}_{1\le l'\le q}\big\}$;
    Compute $\widehat{\mathbf{V}}_{\cdot j}^{(r)} \leftarrow \arg\min_{\{\mathbf{v}:\mathbf{v}_{B^c}=\mathbf{0}\}} \|\mathbf{Y}_{\cdot j} - \mathbf{X}\mathbf{v}\|_2^2$;
  end
  Compute $\widehat{\mathbf{V}}_{\cdot j} \leftarrow \arg\min_{\{\widehat{\mathbf{V}}_{\cdot j}^{(r)}\}_{1\le r\le R}} \|\mathbf{Y}_{\cdot j} - \mathbf{X}\widehat{\mathbf{V}}_{\cdot j}^{(r)}\|_2^2$;
end
return $\widehat{\mathbf{V}} \leftarrow (\widehat{\mathbf{V}}_{\cdot 1}, \ldots, \widehat{\mathbf{V}}_{\cdot p})$;

Remark 6 *In Algorithm 1, γ_j is chosen from a set of candidate values $\{\gamma^{(r)}\}_{1\le r\le R}$. In our implementation, γ_j is not directly tuned by the user and $\{\gamma^{(r)}\}_{1\le r\le R}$ is provided by default. When $\{(\kappa_j, \tau_j)\}_{1\le j\le p}$ are suitably specified by the user, for any value $\gamma^{(r)}$ lies in the proper ranges, $(\mathbf{V}_{\cdot 1}, \ldots, \mathbf{V}_{\cdot p})$ are the global solutions of (7) almost surely; see Theorem 14 in Section 4.1. Moreover, solving the DC programs for $\gamma^{(1)} > \cdots > \gamma^{(R)}$ is efficient with the warm start trick (Friedman et al., 2010; Breheny and Huang, 2011).*

3.1.2 Peeling

Now, we describe a *peeling* algorithm to estimate $\mathcal{G}_+$ based on $\mathbf{V}$. Proposition 5 suggests that the leaf variables of $\mathcal{G}$ (with their instruments) can be identified based on matrix $\mathbf{V}$. To proceed, let $\mathcal{L}$ and its complement $\mathcal{L}^c$ be (generic) nonempty subsets of $\{1, \ldots, p\}$ such that $\boldsymbol{Y}_{\mathcal{L}^c}$ are non-descendants of $\boldsymbol{Y}_{\mathcal{L}}$. Define a sub-DAG $\mathcal{G}_{\mathcal{L}^c} = (\boldsymbol{Y}_{\mathcal{L}^c}, \boldsymbol{X}; \mathcal{E}_{\mathcal{L}^c}, \mathcal{I}_{\mathcal{L}^c})$, where $\mathcal{E}_{\mathcal{L}^c} \subseteq \mathcal{E}$ is the set of primary edges among $\boldsymbol{Y}_{\mathcal{L}^c}$ and $\mathcal{I}_{\mathcal{L}^c} \subseteq \mathcal{I}$ is the set of intervention edges between $\boldsymbol{X}$ and $\boldsymbol{Y}_{\mathcal{L}^c}$. The following proposition offers insights into the connection between $\mathbf{V}$ and $\mathcal{G}_+$.

Proposition 7 *Suppose Assumption 1 is satisfied. Let Y_k be a leaf in $\mathcal{G}_{\mathcal{L}^c}$ and Y_j be in $\boldsymbol{Y}_{\mathcal{L}}$.*

(A) If $\mathrm{V}_{lj} \neq 0$ for each instrument X_l of Y_k in $\mathcal{G}_{\mathcal{L}^c}$, we have $(k, j) \in \mathcal{E}_+$.

(B) If Y_k is an unmediated parent of Y_j, then $\mathrm{V}_{lj} \neq 0$ for each instrument X_l of Y_k in $\mathcal{G}_{\mathcal{L}^c}$.

Proposition 7 (proved in Appendix B.3) together with Proposition 5 indicates that $\mathcal{G}_+$ can be constructed from $\mathbf{V}$. In particular, we can sequentially identify each leaf Y_k with its instrument(s) X_l in the DAG $\mathcal{G}$ or the sub-DAG $\mathcal{G}_{\mathcal{L}^c}$ (where $\boldsymbol{Y}_{\mathcal{L}}$ are peeled variables). Then $\mathcal{E}_+$ can be constructed by including all edges (k, j) such that Y_k is a leaf in the sub-DAG $\mathcal{G}_{\mathcal{L}^c}$, Y_j is a peeled variable, and $\mathrm{V}_{lj} \neq 0$ for each instrument X_l of leaf Y_k in $\mathcal{G}_{\mathcal{L}^c}$. By

Proposition 7, (A) confirms that all such edges are in $\mathcal{E}_+$ so no extra edges are included, and (B) guarantees that every directed edge from an unmediated parent must be included, which is sufficient to determine all the ancestral relations. Thus, $\mathcal{E}_+$ can be recovered from $\mathbf{V}$. Then $\mathcal{I}_+$ is equal to $\{(l,j) : \mathrm{V}_{lk} \neq 0 \text{ if } k = j \text{ or } (k,j) \in \mathcal{E}_+\}$.

The peeling algorithm is summarized in Algorithm 2 and a detailed illustration is presented in Example 3 of Appendix A.2.

Algorithm 2: Reconstruction of ARG by peeling

Input: matrix $\widehat{\mathbf{V}}$.
Output: estimated ARG $\widehat{\mathcal{G}}_+$.
Initialize $\mathbf{V}^{\text{work}} \leftarrow \widehat{\mathbf{V}}$, $\widehat{\mathcal{E}}_+ \leftarrow \emptyset$, $\widehat{\mathcal{I}}_+ \leftarrow \{(l,k) : \widehat{\mathrm{V}}_{lk} \neq 0\}$;
Initialize $\mathcal{G}^{\text{work}}$ by $\mathcal{V}_Y \leftarrow \{1,\ldots,p\}$, $\mathcal{V}_X \leftarrow \{1,\ldots,q\}$, $\mathcal{E}_- \leftarrow \widehat{\mathcal{E}}_+$, $\mathcal{I}_- \leftarrow \widehat{\mathcal{I}}_+$;
while $\mathcal{V}_Y$ *is not empty* **do**
  In $\mathcal{G}^{\text{work}}$, identify the instruments on leaves
  $\mathcal{B} \leftarrow \big\{l : l \text{ minimizes } \|\mathbf{V}^{\text{work}}_{l,\cdot}\|_0 \text{ and } \|\mathbf{V}^{\text{work}}_{l,\cdot}\|_0 > 0\big\}$ and the leaves
  $\mathcal{L} \leftarrow \bigcup_{l \in A} \big\{k : k \text{ maximizes } |\mathrm{V}^{\text{work}}_{lk}|\big\}$;
  Let $\mathcal{B}_k \leftarrow \big\{l \in \mathcal{B} : k \text{ maximizes } |\mathrm{V}^{\text{work}}_{lk}|\big\}$ be the instruments for each $k \in \mathcal{L}$;
  Update $\widehat{\mathcal{E}}_+ \leftarrow \widehat{\mathcal{E}}_+ \cup \big\{(k,j) : j \in \{1,\ldots,p\} \setminus \mathcal{V}_Y,\ k \in \mathcal{L},\ \widehat{\mathrm{V}}_{lj} \neq 0 \text{ for all } l \in \mathcal{B}_k\big\}$;
  Update $\mathcal{G}^{\text{work}}$ by $\mathcal{V}_Y \leftarrow \mathcal{V}_Y \setminus \mathcal{L}$ and update $\mathbf{V}^{\text{work}}$ by keeping the columns in $\mathcal{V}_Y$;
end
Update $\widehat{\mathcal{E}}_+ \leftarrow \big\{(k,j) : Y_k \to \cdots \to Y_j \text{ in } \widehat{\mathcal{E}}_+\big\}$;
Update $\widehat{\mathcal{I}}_+ \leftarrow \widehat{\mathcal{I}}_+ \cup \big\{(l,j) : (l,k) \in \widehat{\mathcal{I}}_+ \text{ and } (k,j) \in \widehat{\mathcal{E}}_+\big\}$;
return $\widehat{\mathcal{E}}_+$ *and* $\widehat{\mathcal{I}}_+$;

Remark 8 *Given* $\widehat{\mathcal{G}}_+ = (\boldsymbol{Y}, \boldsymbol{X}; \widehat{\mathcal{E}}_+, \widehat{\mathcal{I}}_+)$*, we estimate* $(\mathbf{U}_{\widehat{\mathcal{E}}_+}, \mathbf{W}_{\widehat{\mathcal{I}}_+})$ *column-wise from* (5)*,*

$$\min \sum_{i=1}^{n} \Big(Y_{ij} - \sum_{k \in \mathrm{AN}_{\widehat{\mathcal{G}}_+}(j)} \mathrm{U}_{kj} Y_{ik} - \sum_{l \in \mathrm{IN}_{\widehat{\mathcal{G}}_+}(j)} \mathrm{W}_{lj} X_{il} \Big)^2 \text{ s.t. } \sum_{k \in \mathrm{AN}_{\widehat{\mathcal{G}}_+}(j)} \mathrm{I}(\mathrm{U}_{kj} \neq 0) \leq \kappa'_j. \quad (9)$$

The final estimates are $\widehat{\mathbf{U}} = (\widehat{\mathbf{U}}_{\widehat{\mathcal{E}}_+}, \mathbf{0})$ *and* $\widehat{\mathbf{W}} = (\widehat{\mathbf{W}}_{\widehat{\mathcal{I}}_+}, \mathbf{0})$*. In* (9)*, the sparsity constraint is imposed to recover the nonzero elements of* $\mathbf{U}$*; see Appendix A.6 for the technical discussion.*

3.2 Likelihood Inference for a DAG

Now, we propose an inference method for testing (2) and (3). First, we derive the likelihood ratio based on $\mathcal{G}_+$. Next, we perform tests via data perturbation, accounting for the uncertainty of estimating $\mathcal{G}_+$.

3.2.1 Likelihood ratio, nondegeneracy, and regularity

We commence with the likelihood inference for (2), since (3) can be treated as an extension of (2); see the discussion followed by (3). From (5), the maximum likelihood becomes

$$\max_{\mathcal{G}_+, \boldsymbol{\Sigma}} \max_{\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{G}_+}, \mathbf{0})} L(\boldsymbol{\theta}, \boldsymbol{\Sigma}).$$

Thus, we define the likelihood ratio by

$$\mathrm{Lr} = L\Big(\widehat{\boldsymbol{\theta}}^{(1)}, \widehat{\boldsymbol{\Sigma}}\Big) - L\Big(\widehat{\boldsymbol{\theta}}^{(0)}, \widehat{\boldsymbol{\Sigma}}\Big) = \sum_{j=1}^{p} \Big(\mathrm{RSS}_j(\widehat{\boldsymbol{\theta}}^{(0)}) - \mathrm{RSS}_j(\widehat{\boldsymbol{\theta}}^{(1)})\Big)/2\widehat{\sigma}_j^2, \tag{10}$$

where $\widehat{\boldsymbol{\theta}}^{(0)} = (\widehat{\boldsymbol{\theta}}^{(0)}_{\widehat{\mathcal{M}}}, \mathbf{0})$ and $\widehat{\boldsymbol{\theta}}^{(1)} = (\widehat{\boldsymbol{\theta}}^{(1)}_{\widehat{\mathcal{M}}}, \mathbf{0})$ are MLEs (given $\widehat{\mathcal{M}}$) under H_0 and H_a, respectively, $\widehat{\mathcal{M}} = (\boldsymbol{Y}, \boldsymbol{X}; \widehat{\mathcal{E}}_+ \cup \mathcal{H}, \widehat{\mathcal{I}}_+)$ is an estimate for $\mathcal{G}_+$, and $\widehat{\boldsymbol{\Sigma}}$ is an estimate for $\widehat{\boldsymbol{\Sigma}}$. Instead of $\widehat{\mathcal{G}}_+ = (\boldsymbol{Y}, \boldsymbol{X}; \widehat{\mathcal{E}}_+, \widehat{\mathcal{I}}_+)$, the graph $\widehat{\mathcal{M}}$ (with hypothesized edges being added) is used because we need to test the presence of any edge in $\mathcal{H}$.

In many statistical models, the likelihood ratio often has a nondegenerate and tractable distribution when H_0 is true, for instance, a chi-squared distribution with degrees of freedom $|\mathcal{H}|$. However, since $\mathcal{H}$ is pre-specified by the user, $\widehat{\mathcal{M}}$ may not be a DAG, and thus not all edges in $\mathcal{H}$ could present in the DAG parameterized by $(\widehat{\boldsymbol{\theta}}^{(1)}_{\widehat{\mathcal{M}}}, \mathbf{0})$. As a result, Lr for (2) may converge to a distribution with degrees of freedom less than $|\mathcal{H}|$ and the distribution may be even intractable, making inference for a DAG greatly different from the classical ones, as illustrated by Example 1.

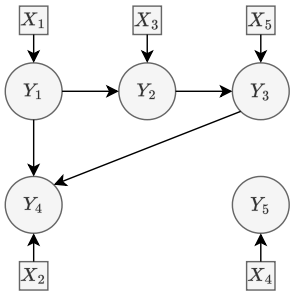


Figure 1: A DAG $\mathcal{G}$ of five primary variables $Y_1, \ldots, Y_5$ and five intervention variables $X_1, \ldots, X_5$, where directed edges are represented by solid arrows while dependencies among $\boldsymbol{X}$ are not displayed.

Example 1 *Consider the likelihood ratio test under null H_0 and alternative H_a for the DAG $\mathcal{G}$ displayed in Figure 1. For simplicity, assume $\widehat{\mathcal{G}}_+ = \mathcal{G}_+$ and $\widehat{\mathcal{M}} = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}_+ \cup \mathcal{H}, \mathcal{I}_+)$.*

- *$H_0 : \mathrm{U}_{21} = 0$ versus $H_a : \mathrm{U}_{21} \neq 0$, where $\mathcal{H} = \{(2,1)\}$. Here, $(2,1)$ forms a cycle together with the edges in $\mathcal{E} \setminus \mathcal{H}$ (namely the edges not considered by the hypothesis), and thus $\widehat{\mathcal{M}}$ has a directed cycle. Given $\widehat{\mathcal{M}}$, when a random sample is obtained under H_0, the likelihood tends to be maximized under the ARG $\mathcal{G}_+$ corresponding to the underlying DAG (which implies $\mathrm{U}_{21} = 0$), especially so when the asymptotics kicks in as the sample size increases. Consequently, the likelihood ratio Lr becomes zero, constituting a degenerate situation.*

- *$H_0 : \mathrm{U}_{45} = \mathrm{U}_{53} = 0$ versus $H_a : \mathrm{U}_{45} \neq 0$ or $\mathrm{U}_{53} \neq 0$, where $\mathcal{H} = \{(4,5),(5,3)\}$. In this case, $\{(4,5),(5,3)\}$ forms a cycle with the edges in $\mathcal{E} \setminus \mathcal{H}$, and $\widehat{\mathcal{M}}$ is cyclic. Given*

$\widehat{\mathcal{M}}$, the likelihood tends to be maximized under either DAG $(\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}_+ \cup \{(4,5)\}, \mathcal{I}_+)$ or DAG $(\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}_+ \cup \{(5,3)\}, \mathcal{I}_+)$ when data is sampled under H_0. Thus, we have

$$L(\widehat{\boldsymbol{\theta}}^{(1)}, \widehat{\boldsymbol{\Sigma}}) = \max\Big(L(\widehat{\mathrm{U}}_{45}, \widehat{\mathrm{U}}_{53} = 0, \widehat{\mathbf{U}}_{\mathcal{H}^c}, \widehat{\mathbf{W}}, \widehat{\boldsymbol{\Sigma}}), L(\widehat{\mathrm{U}}_{45} = 0, \widehat{\mathrm{U}}_{53}, \widehat{\mathbf{U}}_{\mathcal{H}^c}, \widehat{\mathbf{W}}, \widehat{\boldsymbol{\Sigma}})\Big).$$

As a result, the likelihood ratio distribution becomes complicated in this situation due to the dependence between the two components in $L(\widehat{\boldsymbol{\theta}}^{(1)}, \widehat{\boldsymbol{\Sigma}})$.

Motivated by Example 1, we introduce the concepts of nondegeneracy and regularity.

Definition 9 (Nondegeneracy and regularity with respect to DAG)

(A) An edge $(k, j) \in \mathcal{H}$ is nondegenerate with respect to DAG $\mathcal{G}$ if $\{(k, j)\} \cup \mathcal{E}$ contains no directed cycle, where $\mathcal{E}$ denotes the edge set of $\mathcal{G}$. Otherwise, (k, j) is degenerate. Let $\mathcal{D} \subseteq \mathcal{H}$ be the set of all nondegenerate edges with respect to $\mathcal{G}$. A null hypothesis H_0 is nondegenerate with respect to DAG $\mathcal{G}$ if $\mathcal{D} \neq \emptyset$. Otherwise, H_0 is degenerate.

(B) A null hypothesis H_0 is said to be regular with respect to DAG $\mathcal{G}$ if $\mathcal{D} \cup \mathcal{E}$ contains no directed cycle, where $\mathcal{E}$ denotes the edge set of $\mathcal{G}$. Otherwise, H_0 is called irregular.

Remark 10 *In practice, $\mathcal{D}$ is unknown and needs to be estimated from data. Indeed, $\widehat{\mathcal{D}}$ can be computed based on the estimated ARG $\widehat{\mathcal{G}}_+$, because a directed edge (k, j) is nondegenerate if and only if $\{(k, j)\} \cup \mathcal{E}_+$ contains no directed cycle.*

Nondegeneracy ensures nonnegativity of the likelihood ratio. In testing (2), regularity excludes intractable situations for the null distribution. In testing (3), if H_0 is irregular, then $\mathcal{D} \cup \mathcal{E}$ has a directed cycle, which means the hypothesized directed pathway cannot exist due to the acyclicity constraint. Thus, regularity excludes the degenerate situations in testing (3). In what follows, we mainly focus on nondegenerate and regular hypotheses. For the degenerate case, the p-value is defined to be one. For the irregular case of edge test (2), we decompose the hypothesis into regular sub-hypotheses and conduct multiple testing. For the irregular case of pathway test (3), the p-value is defined to be one. More discussions on the implementation in irregular cases are provided in Appendix A.5.

3.2.2 Testing directed edges via data perturbation

Assuming H_0 is nondegenerate and regular, then $\widehat{\boldsymbol{\theta}}^{(1)}$ is the MLE subject to the DAG $\widehat{\mathcal{S}} = (\boldsymbol{Y}, \boldsymbol{X}; \widehat{\mathcal{E}}_+ \cup \widehat{\mathcal{D}}, \widehat{\mathcal{I}}_+)$ and $\widehat{\boldsymbol{\theta}}^{(0)}$ is the MLE subject to an additional constraint $\mathbf{U}_{\mathcal{H}} = \mathbf{0}$. The likelihood ratio (10) can be further simplified. Let $\mathrm{D}_{\widehat{\mathcal{S}}}(j) = \{k : (k, j) \in \widehat{\mathcal{D}}\}$ in DAG $\widehat{\mathcal{S}}$, where $\widehat{\mathcal{D}}$ is the estimated set of nondegenerate edges of H_0 with respect to $\mathcal{G}$. Furthermore, observe that if $\mathrm{D}_{\widehat{\mathcal{S}}}(j) = \emptyset$, then $\mathrm{RSS}_j(\widehat{\boldsymbol{\theta}}^{(0)}) = \mathrm{RSS}_j(\widehat{\boldsymbol{\theta}}^{(1)})$. Hence, Lr only summarizes the contributions from the primary variables with the (estimated) nondegenerate hypothesized edges,

$$\mathrm{Lr} = \sum_{\{j : \mathrm{D}_{\widehat{\mathcal{S}}}(j) \neq \emptyset\}} \Big(\mathrm{RSS}_j(\widehat{\boldsymbol{\theta}}^{(0)}) - \mathrm{RSS}_j(\widehat{\boldsymbol{\theta}}^{(1)})\Big)/2\widehat{\sigma}_j^2, \tag{11}$$

where we estimate $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_j^1, \ldots, \sigma_p^2)$ by

$$\widehat{\sigma}_j^2 = \mathrm{RSS}_j(\widehat{\boldsymbol{\theta}}^{(1)})/\big(n - |\mathrm{PA}_{\widehat{\mathcal{S}}}(j)| - |\mathrm{IN}_{\widehat{\mathcal{S}}}(j)|\big), \quad j = 1, \ldots, p. \tag{12}$$

The likelihood ratio (11) for testing directed edges (2) requires an estimation of $\mathcal{G}_+$ (and $\mathcal{S}$), where we must account for the uncertainty of $\widehat{\mathcal{G}}_+$ (and $\widehat{\mathcal{S}}$) for finite-sample inference. To proceed, we consider the test statistic Lr based on a "correct" ARG $\mathcal{M} \supseteq \mathcal{G}_+$, where $\mathcal{M} = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{A}, \mathcal{C}) \supseteq \mathcal{G}_+ = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}_+, \mathcal{I}_+)$ means that $\mathcal{A} \supseteq \mathcal{E}_+$ and $\mathcal{C} \supseteq \mathcal{I}_+$. Intuitively, a "correct" ARG distinguishes descendants and nondescendants, and thus can help infer the true directed relations defined by the local Markov property (Spirtes et al., 2000) without introducing model errors, yet may lead to a less powerful test when $\mathcal{M}$ is much larger than $\mathcal{G}_+$. By comparison, a "wrong" ARG $\mathcal{M} \not\supseteq \mathcal{G}_+$ may provide an incorrect topological order, and a test based on a "wrong" ARG may be biased, accompanied by an inflated type-I error.

Let $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$ denote the data matrix of primary and intervention variables and let $\mathbf{e} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^\top$ denote the error matrix, where the rows $\{\mathbf{Z}_{i,\cdot} = (\boldsymbol{Y}_i^\top, \boldsymbol{X}_i^\top)\}_{1 \le i \le n}$ and $\{\boldsymbol{\varepsilon}_i^\top\}_{1 \le i \le n}$ are sampled independently from (1). From (11), assuming $\widehat{\mathcal{G}}_+ \supseteq \mathcal{G}_+$ is a "correct" ARG, the likelihood ratio becomes

$$\begin{aligned} \mathrm{Lr} &= \sum_{\{j: \mathrm{D}_{\widehat{\mathcal{S}}}(j) \neq \emptyset\}} \frac{\big\|(\mathbf{P}_{\widehat{A}_j} - \mathbf{P}_{\widehat{B}_j})^{1/2} \mathbf{Y}_{\cdot j}\big\|_2^2}{2\big\|(\mathbf{I} - \mathbf{P}_{\widehat{A}_j})^{1/2} \mathbf{Y}_{\cdot j}\big\|_2^2 / (n - |\widehat{A}_j|)} \\ &= \sum_{\{j: \mathrm{D}_{\widehat{\mathcal{S}}}(j) \neq \emptyset\}} \frac{\big\|(\mathbf{P}_{\widehat{A}_j} - \mathbf{P}_{\widehat{B}_j})^{1/2} (\mathbf{Y}_{\cdot, \mathrm{D}_{\widehat{\mathcal{S}}}(j)} \mathbf{U}_{\mathrm{D}_{\widehat{\mathcal{S}}}(j), j} + \mathbf{e}_{\cdot j})\big\|_2^2}{2\big\|(\mathbf{I} - \mathbf{P}_{\widehat{A}_j})^{1/2} \mathbf{e}_{\cdot j}\big\|_2^2 / (n - |\widehat{A}_j|)}, \end{aligned} \tag{13}$$

where $\mathbf{P}_A = \mathbf{Z}_{\cdot,A} (\mathbf{Z}_{\cdot,A}^\top \mathbf{Z}_{\cdot,A})^{-1} \mathbf{Z}_{\cdot,A}^\top$ is the projection matrix onto the column span of $\mathbf{Z}_{\cdot,A}$, $\widehat{A}_j = \mathrm{PA}_{\widehat{\mathcal{S}}}(j) \cup \mathrm{IN}_{\widehat{\mathcal{S}}}(j)$, and $\widehat{B}_j = \big(\mathrm{PA}_{\widehat{\mathcal{S}}}(j) \cup \mathrm{IN}_{\widehat{\mathcal{S}}}(j)\big) \setminus \mathrm{D}_{\widehat{\mathcal{S}}}(j)$ for $1 \le j \le p$. In (13), we have $\mathbf{Y}_{\cdot, \mathrm{D}_{\mathcal{S}}(j)} \mathbf{U}_{\mathrm{D}_{\mathcal{S}}(j), j} = \mathbf{0}$ for all j under the null hypothesis H_0, while $\mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)} \mathbf{U}_{\mathrm{D}_{\mathcal{S}}(j), j} \neq \mathbf{0}$ for some j under the alternative hypothesis H_a, and thus Lr tends to be large under H_a.

Now, we propose the data perturbation (DP) method (Shen and Ye, 2002; Breiman, 1992) to approximate the null distribution of Lr in (13). The idea behind DP is to assess the sensitivity of the estimates through perturbed data $\mathbf{Y}^* = \mathbf{Y} + \mathbf{e}^*$, where the rows $\{(\boldsymbol{\varepsilon}_i^*)^\top\}_{1 \le i \le n}$ of perturbation errors $\mathbf{e}_{n \times p}^*$ is sampled independently from $N(0, \widehat{\boldsymbol{\Sigma}})$. Let $(\mathbf{Z}^*, \mathbf{e}^*) = (\mathbf{X}, \mathbf{Y}^*, \mathbf{e}^*)$ be the DP sample. Note that the perturbation errors $\mathbf{e}^*$ are only injected into $\mathbf{Y}$ and the perturbation errors $\mathbf{e}^*$ are *known* in the DP sample $(\mathbf{Z}^*, \mathbf{e}^*)$. Given $(\mathbf{Z}^*, \mathbf{e}^*)$, we compute the perturbation estimate $\widehat{\mathcal{G}}_+^*$ (and $\widehat{\mathcal{S}}^*$) by Algorithms 1-2. In (13), under the null hypothesis H_0, the likelihood ratio $\mathrm{Lr} = \Lambda(\mathbf{Z}, \mathbf{e})$ is a function of observed data $\mathbf{Z}$ and unobserved errors $\mathbf{e}$. By definition, the perturbation error $\mathbf{e}^*$ is accessible in the DP sample $(\mathbf{Z}^*, \mathbf{e}^*)$, suggesting the DP likelihood ratio $\mathrm{Lr}^* := \Lambda(\mathbf{Z}^*, \mathbf{e}^*)$ that is equal to

$$\mathrm{Lr}^* = \sum_{\{j: \mathrm{D}_{\widehat{\mathcal{S}}}(j) \neq \emptyset\}} \frac{\big\|(\mathbf{P}^*_{\widehat{A}_j^*} - \mathbf{P}^*_{\widehat{B}_j^*})^{1/2} \mathbf{e}^*_{\cdot j}\big\|_2^2}{2\big\|(\mathbf{I} - \mathbf{P}^*_{\widehat{A}_j^*})^{1/2} \mathbf{e}^*_{\cdot j}\big\|_2^2 / (n - |\widehat{A}_j^*|)}. \tag{14}$$

Note that (14) mimics (13) when $\mathbf{Y}_{\cdot, \mathrm{D}_S(j)} \mathbf{U}_{\mathrm{D}_S(j)} = \mathbf{0}$. As a result, when $\widehat{\mathcal{G}}_+^* \supseteq \mathcal{G}_+$, the conditional distribution of Lr^* given the data $\mathbf{Z}$ well approximates the null distribution of Lr, where the model selection effect is accounted for by assessing the variability of $\{\widehat{A}_j^*, \widehat{B}_j^*\}_{1 \le j \le p}$ over different realizations of $(\mathbf{Z}^*, \mathbf{e}^*)$.

In practice, we use Monte-Carlo to approximate the distribution of Lr^* given $\mathbf{Z}$. We generate M perturbed samples $\{(\mathbf{Z}^*_m, \mathbf{e}^*_m)\}_{1\le m\le M}$ independently and compute $\{\mathrm{Lr}^*_m\}_{1\le m\le M}$, respectively. Then, we examine the condition $\widehat{\mathcal{G}}^*_{+,m} \supseteq \mathcal{G}_+$ by checking its empirical counterpart $\widehat{\mathcal{G}}^*_{+,m} \supseteq \widehat{\mathcal{G}}_+$. The DP p-value of the edge test in (2) is defined as

$$\mathrm{Pval} = \Big(\sum_{m=1}^{M} \mathrm{I}(\mathrm{Lr}^*_m \ge \mathrm{Lr}, \widehat{\mathcal{G}}^*_{+,m} \supseteq \widehat{\mathcal{G}}_+) \Big) / \Big(\sum_{m=1}^{M} \mathrm{I}(\widehat{\mathcal{G}}^*_{+,m} \supseteq \widehat{\mathcal{G}}_+) \Big), \tag{15}$$

where $\mathrm{I}(\cdot)$ is the indicator function.

Remark 11 *Instead of* (14)*, a naive approach is to recompute the likelihood ratio by treating the perturbed sample* $\mathbf{Z}^*$ *as* $\mathbf{Z}$ *while not using the information of* $\mathbf{e}^*$*. However, this is infeasible. For explanation, assuming* $\widehat{\mathcal{G}}^*_+ \supseteq \mathcal{G}_+$ *is a "correct" ARG, then this naive likelihood ratio is equal to*

$$\sum_{\{j:\mathrm{D}_{\widehat{\mathcal{S}}^*}(j)\neq\emptyset\}} \frac{\big\|(\mathbf{P}^*_{\widehat{A}^*_j} - \mathbf{P}^*_{\widehat{B}^*_j})^{1/2}(\mathbf{Y}_{\cdot,j} + \mathbf{e}^*_{\cdot j})\big\|_2^2}{2\big\|(\mathbf{I} - \mathbf{P}^*_{\widehat{A}^*_j})^{1/2}(\mathbf{Y}_{\cdot,j} + \mathbf{e}^*_{\cdot j})\big\|_2^2/(n - |\widehat{A}^*_j|)}.$$

Note that $\{\mathbf{Y}_{\cdot,j}\}_{1\le j\le p}$ *given* $\mathbf{Z}$ *are deterministic and do not vanish under either* H_0 *or* H_a*. Thus, the conditional distribution of this naive likelihood ratio given* $\mathbf{Z}$ *does not approximate the null distribution of Lr, in contrast to the DP likelihood ratio in* (14)*.*

3.2.3 Extension to hypothesis testing for a directed pathway

Next, we extend the DP inference for (3). Denote $\mathcal{H} = \{(k_1, j_1), \ldots, (k_{|\mathcal{H}|}, j_{|\mathcal{H}|})\}$. Then the test of pathways in (3) can be reduced to testing sub-hypotheses

$$H_{0,\nu} : \mathrm{U}_{k_\nu, j_\nu} = 0, \quad \text{versus} \quad H_{a,\nu} : \mathrm{U}_{k_\nu, j_\nu} \neq 0; \quad \nu = 1, \ldots, |\mathcal{H}|,$$

where testing each sub-hypothesis is a directed edge test. Given $(\widehat{\mathcal{S}}, \widehat{\mathbf{\Sigma}})$, the likelihood ratio for $H_{0,\nu}$ is $\mathrm{Lr}_\nu = L(\widehat{\boldsymbol{\theta}}^{(1)}, \widehat{\mathbf{\Sigma}}) - L(\widehat{\boldsymbol{\theta}}^{(0,\nu)}, \widehat{\mathbf{\Sigma}})$, where $\widehat{\boldsymbol{\theta}}^{(0,\nu)}$ is the MLE under the constraint that $\mathrm{U}_{k_\nu, j_\nu} = 0$. When $\widehat{\mathcal{G}}_+ \supseteq \mathcal{G}_+$, we have

$$\mathrm{Lr}_\nu = \frac{\big\|(\mathbf{P}_{\widehat{A}_{j_\nu}} - \mathbf{P}_{\widehat{B}_{j_\nu}})^{1/2}(\mathbf{Y}_{k_\nu}\mathrm{U}_{k_\nu,j_\nu} + \mathbf{e}_{\cdot j_\nu})\big\|_2^2}{2\big\|(\mathbf{I} - \mathbf{P}_{\widehat{A}_{j_\nu}})^{1/2}\mathbf{e}_{\cdot j_\nu}\big\|_2^2/(n - |\widehat{A}_{j_\nu}|)}, \ \mathrm{Lr}^*_\nu = \frac{\big\|(\mathbf{P}^*_{\widehat{A}^*_{j_\nu}} - \mathbf{P}^*_{\widehat{B}^*_{j_\nu}})^{1/2}\mathbf{e}^*_{\cdot j_\nu}\big\|_2^2}{2\big\|(\mathbf{I} - \mathbf{P}^*_{\widehat{A}^*_{j_\nu}})^{1/2}\mathbf{e}^*_{\cdot j_\nu}\big\|_2^2/(n - |\widehat{A}^*_{j_\nu}|)}, \tag{16}$$

where the distributions of Lr^*_ν given $\mathbf{Z}$ approximates the null distributions of Lr_ν. Finally, define the p-value of the pathway test in (3) as

$$\mathrm{Pval} = \max_{1\le\nu\le|\mathcal{H}|} \Big(\sum_{m=1}^{M} \mathrm{I}(\mathrm{Lr}^*_{\nu,m} \ge \mathrm{Lr}_\nu, \widehat{S}^*_m \supseteq \widehat{S}) \Big) / \Big(\sum_{i=1}^{M} \mathrm{I}(\widehat{S}^*_m \supseteq \widehat{S}) \Big). \tag{17}$$

Note that if any $H_{0,\nu}$ is degenerate, then $\mathrm{Pval} = 1$.

Algorithm 3 summarizes the DP method for hypothesis testing.

Algorithm 3: DP likelihood ratio test

Input: hypothesis H_0; data $\mathbf{Z}$

Parameters: Monte Carlo size M; parameters for Algorithm 1.

Output: the p-value Pval.

Compute $(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_p^2)$ and Lr with data $\mathbf{Z}$;

for *each* $1 \le m \le M$ *in parallel* **do**

  Generate perturbed data $(\mathbf{Z}_m^*, \mathbf{e}_m^*)$;

  For (2), compute Lr_m^* based on (14);

  For (3), compute $\{\mathrm{Lr}_{\nu,m}^*\}_{1 \le \nu \le |\mathcal{H}|}$ based on (16);

end

Compute Pval as (15) or (17) accordingly;

return Pval;

Remark 12 *For acceleration, we parallelize Step 2 in Algorithm 3. Additionally, we use the estimate $\widehat{\boldsymbol{\theta}}$ as a warm-start initialization for the DP estimates, effectively reducing the computing time.*

Remark 13 (Connection with bootstrap) *One may consider parametric or nonparametric bootstrap for Lr. The parametric bootstrap requires a good initial estimate of $(\mathbf{U}, \mathbf{W})$. Yet, it is rather challenging to correct the bias of this estimate because of the acyclicity constraint. By comparison, DP does not rely on such an estimate. On the other hand, nonparametric bootstrap resamples the original data with replacement. In a bootstrap sample, only about 63% distinct observations in the original data are used in model selection and fitting, leading to deteriorating performance (Kleiner et al., 2012), especially in a small sample. As a result, nonparametric bootstrap may not well-approximate the distribution of Lr, while DP provides a better approximation of Lr, taking advantage of a full sample.*

4. Theory

This section provides the theoretical justification for the proposed methods.

4.1 Convergence and Consistency of Structure Learning

First, we introduce some technical assumptions to derive statistical and computational properties of Algorithms 1 and 2. Let $\boldsymbol{\zeta}$ be a generic vector and $\boldsymbol{\zeta}_A$ be the subvector of $\boldsymbol{\zeta}$ with coordinates in A. Let $\kappa_j^\circ = \|\mathbf{V}_{\cdot j}\|_0$ and $\kappa_{\max}^\circ = \max_{1 \le j \le p} \kappa_j^\circ$.

Assumption 2 *For constants $c_1, c_2 > 0$,*

(A) $\min_{\{A:|A| \le 2\kappa_{\max}^\circ\}} \min_{\{\boldsymbol{\zeta}:\|\boldsymbol{\zeta}_{A^c}\|_1 \le 3\|\boldsymbol{\zeta}_A\|_1\}} \|\mathbf{X}\boldsymbol{\zeta}\|_2^2/n \|\boldsymbol{\zeta}\|_2^2 \ge c_1$ *almost surely.*

(B) $\max_{1 \le l \le q} (\mathbf{X}^\top \mathbf{X})_{ll}/n \le c_2^2$ *almost surely.*

Assumption 3 $\min_{\mathrm{V}_{lj} \ne 0} |\mathrm{V}_{lj}| \ge 100 c_1^{-1} c_2 \max_{1 \le j \le p} \left(\Omega_{jj}^{-1/2}\right) \sqrt{\log(q)/n + \log(n)/n}$.

Assumption 2 is a common condition for proving the convergence rate of the Lasso (Bickel et al., 2009; Zhang et al., 2014). As a replacement of Assumption 1A, it can be satisfied for isotropic subgaussian or bounded $\boldsymbol{X}$ (Rudelson and Zhou, 2013). Assumption 3, as an alternative to Assumption 1B, specifies the minimal signal strength over candidate interventions. Such a signal strength requirement is used for establishing the high-dimensional variable selection consistency (Fan et al., 2014; Loh and Wainwright, 2017; Zhao et al., 2018). Moreover, Assumption 3 can be further relaxed to a less intuitive condition, Assumption 5; see Appendix A.3 for details.

Theorem 14 *Suppose Assumptions 1-3 are met with constants $c_1 < 6c_2$, and the machine precision* tol $\ll 1/n$ *is negligible. For $1 \le j \le p$, if the tuning parameters (κ_j, τ_j) of Algorithm 1 are suitably chosen such that*

$$\kappa_j = \kappa_j^\circ, \qquad \frac{36c_2}{c_1}\sqrt{\Omega_{jj}^{-1}\Big(\frac{\log(q)}{n} + \frac{\log(n)}{n}\Big)} \le \tau_j \le \frac{2}{5}\min_{\mathrm{V}_{lj}\neq 0}|\mathrm{V}_{lj}|,$$

then for any γ_j such that $\tau_j^{-1}(32c_2^2\Omega_{jj}^{-1}n^{-1}(\log(q)+\log(n)))^{1/2} \le \gamma_j \le c_1/6$, almost surely we have Algorithm 1 yields a global solution $\widehat{\mathbf{V}}_{\cdot j}$ of (7) in at most $1+\lceil\log(\kappa_{\max}^\circ)/\log(4)\rceil$ DC iterations when n is sufficiently large, where $\lceil\cdot\rceil$ is the ceiling function. Moreover, almost surely we have Algorithm 2 recovers $\mathcal{E}_+$ and $\mathcal{I}_+$ when n is sufficiently large.

In view of Theorem 14 (proved in Appendix B.4), it suffices to specify the maximum number of DC iterations as $1+\lceil\log(\kappa_{\max}^\circ)/\log(4)\rceil$. Then the time complexity of Algorithm 1 is $p\times O(\log\kappa_{\max}^\circ)\times O(q^3+nq^2)$, where $O(q^3+nq^2)$ is that of solving a weighted Lasso (Efron et al., 2004). Note that Algorithm 2 does not involve heavy computation, so the overall time complexity for estimating the ARG (Algorithms 1-2) is $O(p\times\log\kappa_{\max}^\circ\times(q^3+nq^2))$. Finally, the peeling method does not apply to observational data ($\mathbf{W}=\mathbf{0}$). In a sense, interventions are essential.

Theorem 14 establishes the consistent reconstruction by the peeling algorithm for the ARG. Yet, it does not provide any uncertainty measure for the presence of some directed edges in the true DAG. In what follows, we will develop an asymptotic theory for hypothesis tests concerning directed edges of interest.

4.2 Inferential Theory

Given H_0, let $\mathcal{S} = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}_+ \cup \mathcal{D}, \mathcal{I}_+)$.

Assumption 4 $\max_{\{j:\mathrm{D}_{\mathcal{S}}(j)\neq\emptyset\}}\big(|\mathrm{PA}_{\mathcal{S}}(j)| + |\mathrm{IN}_{\mathcal{S}}(j)|\big)/n \le \rho$ *as $n\to\infty$, for a constant $0<\rho<1$.*

Assumption 4 is a hypothesis-specific condition restricting the underlying dimension of the testing problem. Usually, $|\mathrm{PA}_{\mathcal{S}}(j)| \asymp |\mathrm{AN}_{\mathcal{G}}(j)| \asymp |\mathrm{IN}_{\mathcal{S}}(j)| \asymp \kappa_j^\circ \ll p$; $1\le j\le p$, which relaxes the condition $n \gg p\log(p)\sqrt{|\mathcal{D}|}$ for the constrained likelihood ratio test (Li et al., 2020).

Theorem 15 (Empirical p-values) *Suppose Assumptions 1-4 are met and H_0 is regular. Assume the tuning parameters in Algorithm 1 satisfy the requirements in Theorem 14.*

(A) For the test of directed edges (2)*,*

$$\lim_{\substack{n\to\infty \\ \boldsymbol{\theta} \text{ satisfies } H_0 \text{ in (2)}}} \lim_{M\to\infty} \mathbb{P}_{\boldsymbol{\theta}}(\text{Pval} < \alpha) = \alpha, \text{ if } H_0 \text{ is nondegenerate.}$$

$$\lim_{\substack{n\to\infty \\ \boldsymbol{\theta} \text{ satisfies } H_0 \text{ in (2)}}} \lim_{M\to\infty} \mathbb{P}_{\boldsymbol{\theta}}(\text{Pval} = 1) = 1, \text{ if } H_0 \text{ is degenerate.}$$

(B) For the test of directed pathways (3)*,*

$$\limsup_{\substack{n\to\infty \\ \boldsymbol{\theta} \text{ satisfies } H_0 \text{ in (3)}}} \lim_{M\to\infty} \mathbb{P}_{\boldsymbol{\theta}}(\text{Pval} < \alpha) = \alpha, \text{ if } H_0 \text{ is nondegenerate with } |\mathcal{D}| = |\mathcal{H}|.$$

$$\limsup_{\substack{n\to\infty \\ \boldsymbol{\theta} \text{ satisfies } \mathrm{H}_0 \text{ in (3)}}} \lim_{M\to\infty} \mathbb{P}_{\boldsymbol{\theta}}(\text{Pval} = 1) = 1, \text{ if } |\mathcal{D}| < |\mathcal{H}|.$$

By Theorem 15 (proved in Appendix B.5), the DP likelihood ratio test yields a valid p-value for (2) and (3) under appropriate conditions. Note that $|\mathcal{D}|$ is permitted to depend on n. Moreover, Proposition 16 (proved in Appendix B.6) summarizes the asymptotics for directed edge test (2).

Proposition 16 (Asymptotics of edge test) *Suppose the assumptions of Theorem 15 are met. Under a nondegenerate and regular H_0, as $n \to \infty$,*

(A) $2\text{Lr} \xrightarrow{d} \chi^2_{|\mathcal{D}|}$*, if $|\mathcal{D}| > 0$ is fixed.*

(B) $(2\text{Lr} - |\mathcal{D}|)/\sqrt{2|\mathcal{D}|} \xrightarrow{d} N(0,1)$*, if $|\mathcal{D}|\log(|\mathcal{D}|)/n \to 0$.*

Remark 17 *As opposed to the entire $\mathcal{G}_+$ (or $\mathcal{S}$), Theorem 15 and Proposition 16 only require correct identification of the local structures $\left\{\text{AN}_{\mathcal{G}_+}(j), \text{IN}_{\mathcal{G}_+}(j), \text{D}_{\mathcal{S}}(j)\right\}_{\{j:\text{D}_{\mathcal{S}}(j)\neq\emptyset\}}$. This requirement can be reasonably satisfied when the sample size is moderately large, as illustrated in Section 5.*

Next, we analyze the local limit power of the proposed tests for (2) and (3). Assume $\boldsymbol{\theta}^\circ = (\mathbf{U}^\circ, \mathbf{W}^\circ)$ satisfies H_0. Let $\boldsymbol{\Delta} \in \mathbb{R}^{p\times p}$ satisfy $\boldsymbol{\Delta}_{\mathcal{D}^c} = \mathbf{0}$ so that $\mathbf{U}^\circ + \boldsymbol{\Delta}$ represents a DAG. For nondegenerate and regular H_0, consider an alternative H_a: $\mathbf{U}_{\mathcal{H}} = \mathbf{U}^\circ_{\mathcal{H}} + \boldsymbol{\Delta}_{\mathcal{H}}$, and define the power function as

$$\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta}) = \mathbb{P}_{H_a}(\text{Pval} < \alpha). \tag{18}$$

Proposition 18 (Local power of edge test) *Suppose H_0 is nondegenerate and regular. Let $\|\boldsymbol{\Delta}\|_F = \|\boldsymbol{\Delta}_{\mathcal{H}}\|_F = n^{-1/2}\delta$ when $|\mathcal{D}| > 0$ is fixed and $\|\boldsymbol{\Delta}\|_F = |\mathcal{D}|^{1/4}n^{-1/2}h$ when $|\mathcal{D}| \to \infty$, where $\delta > 0$ and $\|\cdot\|_F$ is the matrix Frobenius norm. If the assumptions of Theorem 15 are met, then under H_a, as $n, M \to \infty$,*

$$\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta}) \geq \begin{cases} \mathbb{P}_{\boldsymbol{Z}\sim N(\mathbf{0}, \mathbf{I}_{|\mathcal{D}|\times|\mathcal{D}|})}\left(\|\boldsymbol{Z} + c_l\sqrt{n}\boldsymbol{\Delta}\|_2^2 > \chi^2_{|\mathcal{D}|,1-\alpha}\right) & \text{if } |\mathcal{D}| > 0 \text{ is fixed,} \\ \mathbb{P}_{Z\sim N(0,1)}\left(Z > z_{1-\alpha} - c_l\|\boldsymbol{\Delta}\|_2^2/\sqrt{2|\mathcal{D}|}\right) & \text{if } |\mathcal{D}| \to \infty, \frac{|\mathcal{D}|\log|\mathcal{D}|}{n} \to 0, \end{cases}$$

where $\chi^2_{|\mathcal{D}|,1-\alpha}$ and $z_{1-\alpha}$ denote the $(1-\alpha)$th quantile of distributions $\chi^2_{|\mathcal{D}|}$ and $N(0,1)$, respectively. Hence, $\lim_{\delta\to\infty}\lim_{n\to\infty}\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta}) = 1$.

Proposition 19 (Local power of pathway test) *Suppose H_0 is nondegenerate and regular with $|\mathcal{D}| = |\mathcal{H}|$. Let $\min_{(k,j)\in\mathcal{H}} |\mathrm{U}^\circ_{kj} + \Delta_{kj}| = n^{-1/2}\delta$ when $|\mathcal{H}| > 0$ is fixed and $\min_{(k,j)\in\mathcal{H}} |\mathrm{U}^\circ_{kj} + \Delta_{kj}| = n^{-1/2}\delta\sqrt{\log|\mathcal{H}|}$ when $|\mathcal{H}| \to \infty$. If the assumptions of Theorem 15 are met, then under H_a, as $n, M \to \infty$,*

$$\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta}) \geq 1 - \frac{|\mathcal{H}|}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\Big(\delta\sqrt{\log|\mathcal{H}|}/\max_{1\leq j\leq p}\Omega_{jj} - \sqrt{\chi^2_{1,1-\alpha}}\Big)^2\Big),$$

where $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$th quantile of distribution χ^2_1. Hence, $\lim_{\delta\to\infty}\lim_{n\to\infty}\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta}) = 1$.

The proofs of Propositions 18 and 19 are deferred to Appendix B.7 and B.8.

5. Simulations

This section investigates the operating characteristics of the proposed tests and the peeling algorithm via simulations. In simulations, we consider two setups for generating $\mathbf{U} \in \mathbb{R}^{p\times p}$, representing random and hub DAGs, respectively.

- **Random graph.** The upper off-diagonal entries U_{kj}; $k < j$ are sampled independently from $\{0, 1\}$ according to Bernoulli$(1/p)$, while other entries are zero. This generates a random graph with a sparse neighborhood.

- **Hub graph.** Set $\mathrm{U}_{1,2j+1} = 1$ and $\mathrm{U}_{2,2j+2} = 1$ for $j = 1, \ldots, \lfloor p/2\rfloor - 2$, while other entries are zero. This generates a hub graph, where nodes 1 and 2 are hub nodes with a dense neighborhood.

Moreover, we consider three setups for intervention matrix $\mathbf{W} \in \mathbb{R}^{q\times p}$, representing different scenarios. Setups (A) and (B) are designed for inference, whereas Setup (C) in Section 5.2 is designed to compare with the method of Chen et al. (2018) for structure learning. Let $\mathbf{W} = (\mathbf{A}^\top, \mathbf{B}^\top, \mathbf{0}^\top)^\top$, where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ and $\mathbf{0} \in \mathbb{R}^{(q-2p)\times p}$.

- **Setup (A).** Set $\mathrm{A}_{jj} = \mathrm{B}_{jj} = \mathrm{B}_{j,j+1} = 1$; $j = 1, \ldots, p-1$, $\mathrm{A}_{pp} = 1$, while other entries of $\mathbf{A}$, $\mathbf{B}$ are zero. Then, $X_1, \ldots, X_p$ are instruments for $Y_1, \ldots, Y_p$, respectively, $X_{p+1}, \ldots, X_{2p-1}$ are invalid instruments with two targets, and $X_{2p}, \ldots, X_q$ represent inactive interventions.

- **Setup (B).** Set $\mathrm{A}_{jj} = \mathrm{A}_{j,j+1} = \mathrm{B}_{jj} = \mathrm{B}_{j,j+1} = 1$; $j = 1, \ldots, p-1$, $\mathrm{A}_{pp} = 1$, while other entries of $\mathbf{A}$, $\mathbf{B}$ are zero. Here, the only valid instrument is X_p on Y_p, and the other intervention variables either have two targets or are inactive. Importantly, Assumption 1C is not met.

To generate $(\boldsymbol{Y}, \boldsymbol{X})$ for each setup, we sample $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X)$ with $(\Sigma_X)_{ll'} = 0.5^{|l-l'|}$; $1 \leq l, l' \leq q$ and sample $\boldsymbol{Y}$ according to (1) with $(\mathbf{U}, \mathbf{W}, \sigma_1^2, \ldots, \sigma_p^2)$, where $\sigma_1^2, \ldots, \sigma_p^2$ are set to be equally spaced from 0.5 to 1.

5.1 Inference

We compare three tests in empirical type-I errors and powers in simulated examples, namely, the DP likelihood ratio test (DP-LR) in Algorithm 3, the asymptotic likelihood ratio test (LR), and the oracle likelihood ratio test (OLR). Here LR uses Lr, while OLR uses $\mathrm{Lr}(\mathcal{S}, \widehat{\mathbf{\Sigma}})$ assuming that $\mathcal{S}$ were known in advance. The p-values of LR and OLR are computed via Proposition 16. The implementation details of these tests are in Appendix C.1.

For the empirical type-I error of a test, we compute the percentage of times rejecting H_0 out of 500 simulations when H_0 is true. For the empirical power of a test, we report the percentage of times rejecting H_0 out of 100 simulations when H_a is true under alternative hypotheses H_a.

- **Test of directed edges.** For (2), we examine two different hypotheses:

  (i) $H_0 : \mathrm{U}_{1,20} = 0$ versus $H_a : \mathrm{U}_{1,20} \neq 0$. In this case, $|\mathcal{D}| = 1$.

  (ii) $H_0 : \mathbf{U}_{\mathcal{H}} = \mathbf{0}$ versus $H_a : \mathbf{U}_{\mathcal{H}} \neq \mathbf{0}$, where $\mathcal{H} = \{(k, 20) : k = 1, \ldots, 15\}$. In this case, $|\mathcal{D}| = 15$.

  Moreover, five alternatives $H_a : \mathrm{U}_{1,20} = 0.1l$ and $\mathbf{U}_{\mathcal{H}\setminus\{(1,20)\}} = \mathbf{0}$; $l = 1, 2, 3, 4, 5$, are used for the power analysis in (i) and (ii). The data are generated by modifying $\mathbf{U}$ accordingly.

- **Test of directed pathways.** We test the directed path $Y_1 \to Y_5 \to Y_{10} \to Y_{15} \to Y_{20}$, namely $\mathcal{H} = \{(1, 5), (5, 10), (10, 15), (15, 20)\}$ in (3). Since (3) is a test of composite null hypothesis, the data are generated under a graph with parameters $(\mathbf{U}, \mathbf{W}, \mathbf{\Sigma})$ satisfying H_0, where $\mathbf{U}_{\mathcal{H}} = \mathbf{0}$. Five hypotheses $H_a : \mathbf{U}_{\mathcal{H}} = \mathbf{0.1}l$; $l = 1, 2, 3, 4, 5$ are used for the power analysis.

For testing directed edges, as displayed in Figure 2, DP-LR and LR perform well compared to the ideal test OLR in Setup (A) with Assumption 1 satisfied. In Setup (B) with Assumption 1C not fulfilled, DP-LR appears to have control of type-I error, whereas LR has an inflated empirical type-I error compared to the nominal level $\alpha = 0.05$. This discrepancy is attributed to the data perturbation scheme accounting for the uncertainty of identifying S. However, both suffer a loss of power compared to the oracle test OLR in this setup. This observation suggests that without Assumption 1C, the peeling algorithm tends to yield an estimate $\widehat{\mathcal{G}}_+ \supseteq \mathcal{G}_+$, which overestimates $\mathcal{G}_+$, resulting in a power loss.

For testing directed pathways, as indicated in Figure 3, we observe similar phenomena as in the previous directed edge tests. Of note, both LR and DP-LR are capable in controlling type-I error of directed path tests.

In summary, DP-LR has a suitable control of type-I error when there are invalid instruments and Assumption 1C is violated. Concerning the power, DP-LR and LR are comparable in all scenarios and their powers tend to one as the sample size n or the signal strength of tested edges increases. Moreover, DP-LR and LR perform nearly as well as the oracle test OLR when Assumption 1 is satisfied. These empirical findings agree with our theoretical results.

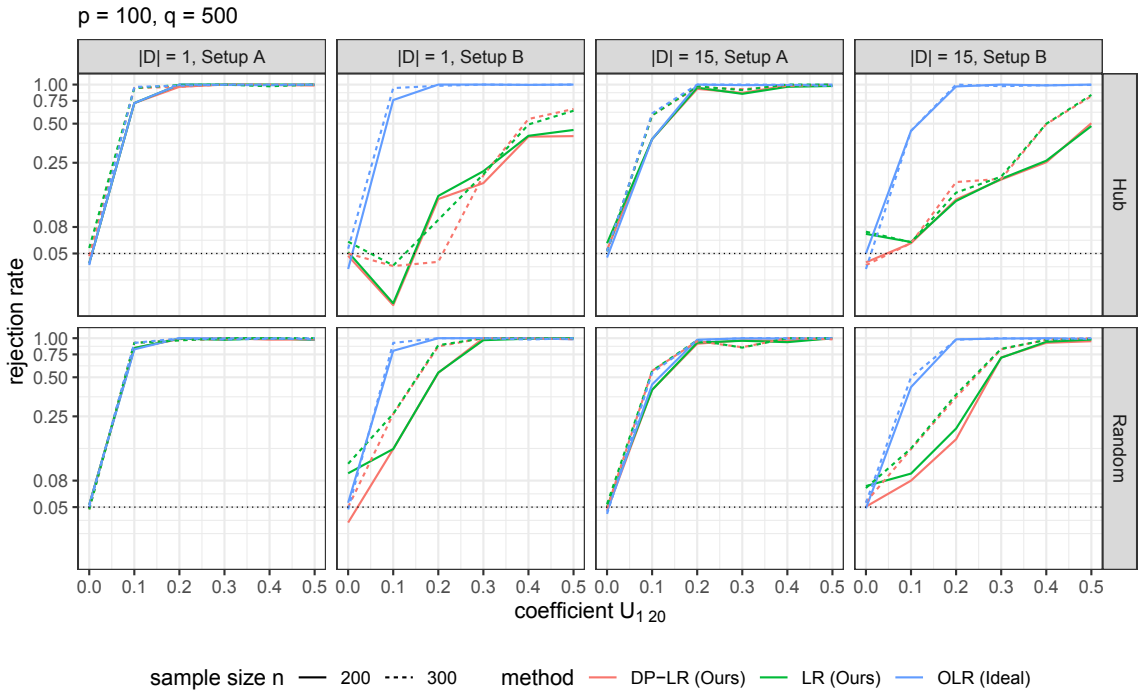


Figure 2: Empirical type-I errors and powers of tests of directed edges. The black dotted line marks the nominal level of significance $\alpha = 0.05$.

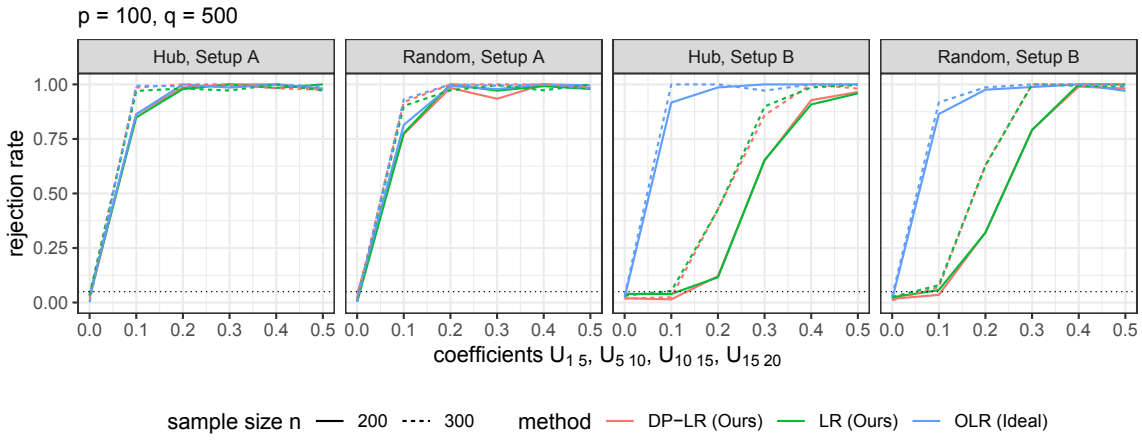


Figure 3: Empirical type-I errors and powers of tests of a directed pathway. The black dotted line marks the nominal level of significance $\alpha = 0.05$.

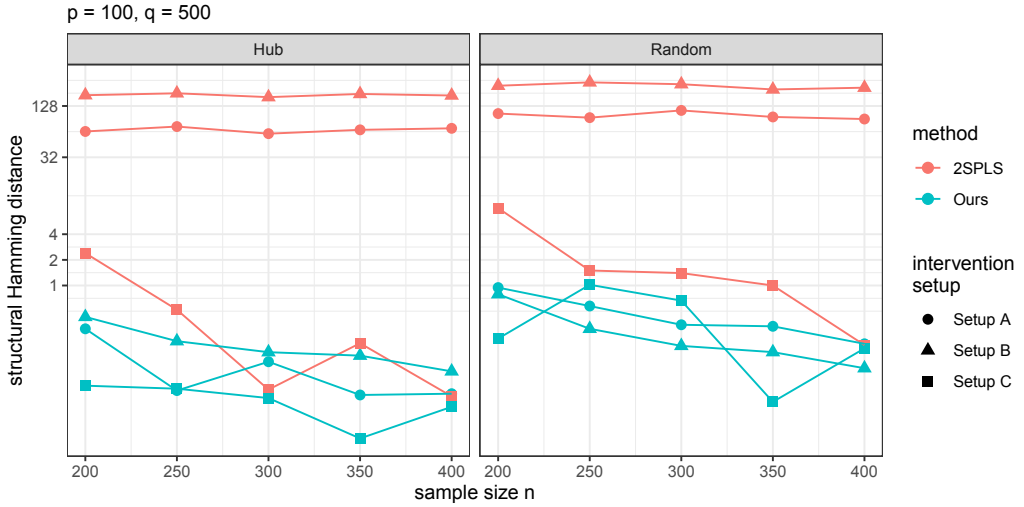


Figure 4: SHDs for the reconstructed DAG by the peeling algorithm and 2SPLS, where a smaller value of SHD indicates a better result.

5.2 Structure Learning

This subsection compares the peeling algorithm with the Two-Stage Penalized Least Squares (2SPLS, Chen et al. (2018)) in terms of the structure learning accuracy. For peeling, we consider Algorithm 2 with an additional step (9) for structure learning of $\mathbf{U}$. For 2SPLS, we use the R package `BigSEM`.

2SPLS requires that all the intervention variables to be target-known instruments in addition to Assumption 1C. Thus, we consider an additional Setup (C).

- **Setup (C).** Let $\mathbf{W} = (\mathbf{I}_{p\times p}, \mathbf{0})^\top \in \mathbb{R}^{q\times p}$. Then $X_1, \ldots, X_p$ are valid instruments for $Y_1, \ldots, Y_p$, respectively, and other intervention variables are inactive.

For 2SPLS, we assign each active intervention variable to its most correlated primary variable in Setups (A)-(C). In Setup (C), this assignment yields a correct identification of valid instruments, meeting all the requirements of 2SPLS.

For each scenario, we compute the structural Hamming distance (SHD)

$$\mathrm{SHD}(\widehat{\mathbf{U}}, \mathbf{U}) = \sum_{k,j} |\,\mathrm{I}(\widehat{\mathrm{U}}_{kj} \neq 0) - \mathrm{I}(\mathrm{U}_{kj} \neq 0)|,$$

averaged over 100 runs. As shown in Figure 4, the peeling algorithm outperforms 2SPLS, especially when there are invalid instruments and Assumption 1C is violated.

Appendix C.2 contains additional numerical experiments on structure learning, including the results of different sparsity settings, SHD transition curves, and different numbers of interventions.

5.3 Comparison of Inference and Structure Learning

This subsection compares the proposed DP testing method against the proposed structure learning method in (9) in terms of inferring the true graph structure. To this end, we consider Setup (A) in Section 5.1 with $p = 30, q = 100$, and the hypotheses

$$H_0 : \mathrm{U}_{1,20} = 0 \text{ versus } H_a : \mathrm{U}_{1,20} = 1/\sqrt{n}.$$

For DP inference, we use $\alpha = 0.05$ and choose the tuning parameters by BIC as in previous experiments; see Appendix C.1 for details. For structure learning, we reject the null hypothesis when $\widehat{\mathrm{U}}_{1,20} \neq 0$.

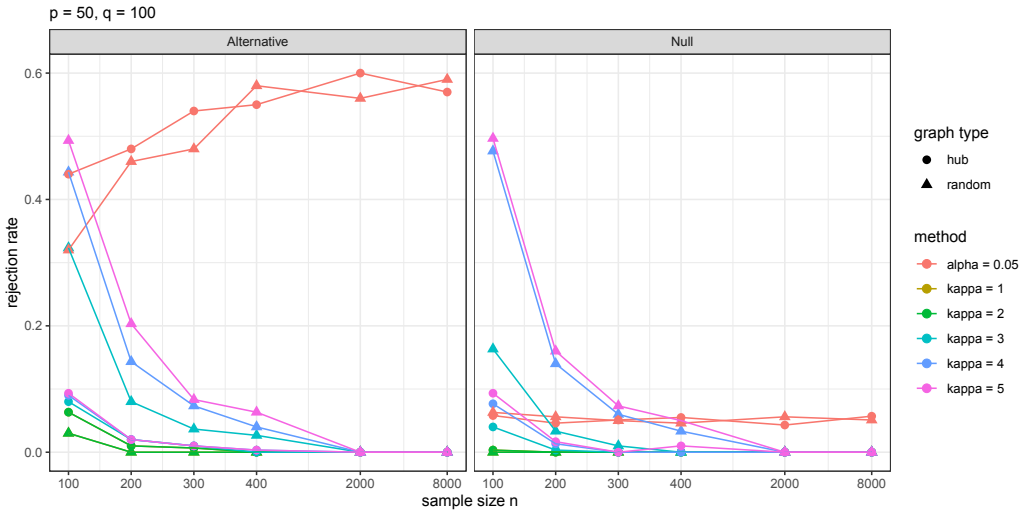


Figure 5: Rejection rates for $H_0 : \mathrm{U}_{1,20} = 0$ versus $H_a : \mathrm{U}_{1,20} = 1/\sqrt{n}$ by DP inference and structure learning. For structure learning, H_0 is rejected if $\widehat{\mathrm{U}}_{1,20} \neq 0$. The red lines indicate the results of DP inference using the significance level $\alpha = 0.05$. The other colored lines display the results of structure learning using different sparsity parameter values $\kappa = 1, 2, 3, 4, 5$. The simulation is repeated for 500 times and $\kappa = 2$ is chosen by BIC in over 90% cases.

As displayed in Figure 5, when the null hypothesis H_0 is true, the DP testing method controls type-I error very close to the nominal level of 0.05, whereas the type-I error of the structure learning varies greatly depending on the tuning parameter selection. Under the alternative hypothesis H_a, the DP inference enjoys high statistical power than structure learning methods when $n \geq 200$. Interestingly, the power of structure learning diminishes as n increases. This observation is in agreement with our theoretical results in Theorem 21, suggesting that consistent reconstruction requires the smallest size of nonzero coefficients to be of order $\gtrsim \sqrt{\log(n)/n}$ with the tuning parameter τ of the same order (fixing p, q). In this case, the edge $\mathrm{U}_{1,20}$ is of order $1/\sqrt{n}$, which is less likely to be reconstructed as n increases. In contrast, Proposition 18 indicates that a DP test has a non-vanishing power when the hypothesized edges are of order $1/\sqrt{n}$.

Figure 5 demonstrates some important distinctions between inference and structure learning. When different tuning parameters are used, the structure learning results correspond to different points on an ROC curve. Although it is asymptotically consistent when optimal tuning parameters are used, structure learning lacks an uncertainty measure of graph structure identification. As a result, it is nontrivial for structure learning methods to trade-off the false discovery rate and detection power in practice. This makes the interpretation of such results hard, especially when they heavily rely on hyperparameters as in Figure 5. By comparsion, DP inference aims to maximize statistical power while controlling type-I error at a given level, offering a clear interpretation of its result. This observation agrees with the discussions in the literature on variable selection and inference (Wasserman and Roeder, 2009; Meinshausen and Bühlmann, 2010; Lockhart et al., 2014; Candes et al., 2018) and it justifies the demand for inferential tools for directed graphical models.

6. ADNI Data Analysis

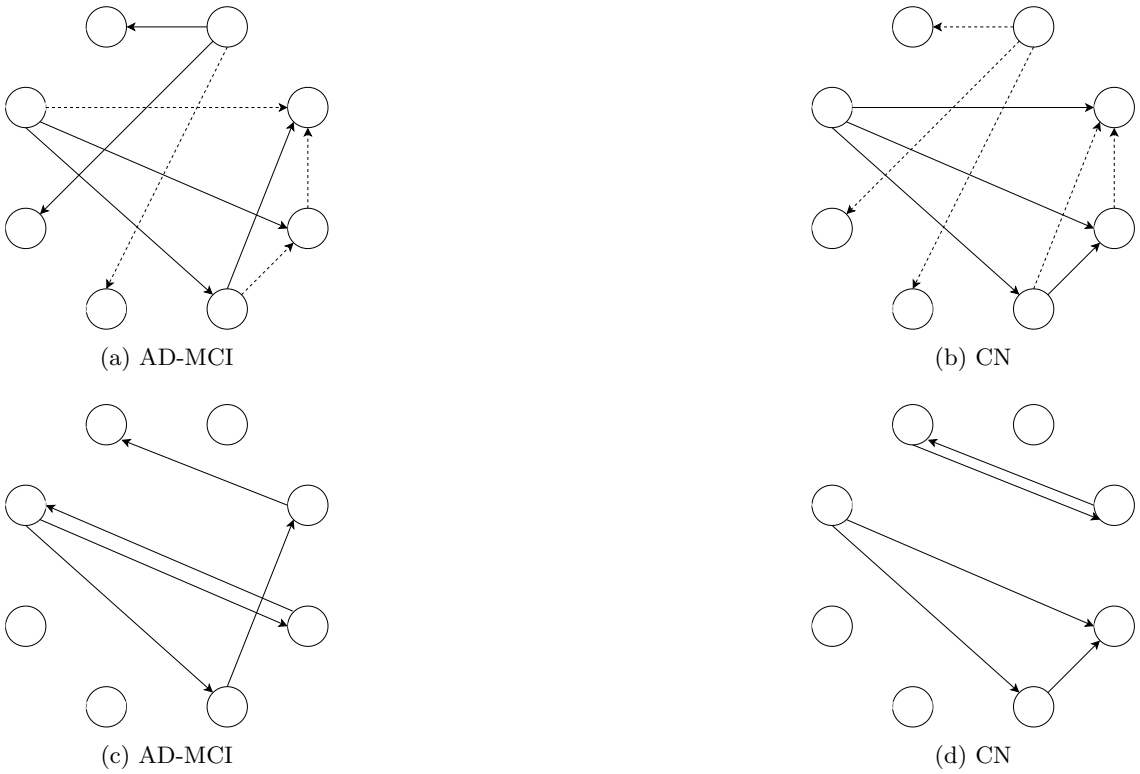
This section applies the proposed tests to analyze an Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. In particular, we infer gene pathways related to Alzheimer's Disease (AD) to highlight some gene-gene interactions differentiating patients with AD/cognitive impairments and healthy individuals.

The raw data are available in the ADNI database (`https://adni.loni.usc.edu`), including gene expression, whole-genome sequencing, and phenotypic data. After cleaning and merging, we have a sample size of 712 subjects. From the KEGG database (Kanehisa and Goto, 2000), we extract the AD reference pathway (hsa05010, `https://www.genome.jp/pathway/hsa05010`), including 146 genes in the ADNI data.

For data analysis, we first regress the gene expression levels on five covariates – gender, handedness, education level, age, and intracranial volume, and then use the residuals as gene expressions in the following analysis. Next, we extract the genes with at least one SNP at a marginal significance level below 10^{-3}, yielding $p = 63$ genes as primary variables. For these genes, we further extract their marginally most correlated two SNPs, resulting in $q = 63 \times 2 = 126$ SNPs as unspecified intervention variables for subsequent data analysis. All gene expression levels are normalized.

The dataset contains individuals in four groups, namely, Alzheimer's Disease (AD), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Cognitive Normal (CN). For our purpose, we treat 247 CN individuals as controls while the remaining 465 individuals as cases (AD-MCI). Then, we use the gene expressions and the SNPs to reconstruct the ancestral relations and infer gene pathways for 465 AD-MCI and 247 CN control cases, respectively.

In the literature, genes APP, CASP3, and PSEN1 are well-known to be associated with AD, reported to play different roles in AD patients and healthy subjects (Julia and Goate, 2017; Su et al., 2001; Kelleher III and Shen, 2017). For this dataset, we conduct hypothesis testing on edges and pathways related to genes APP, CASP3, and PSEN1 in the KEGG AD reference (hsa05010) to evaluate the proposed DP inference by checking if DP inference can discover the differences that are reported in the biomedical literature. First, we consider testing $H_0 : \mathrm{U}_{kj} = 0$ versus $H_a : \mathrm{U}_{kj} \neq 0$, for each edge (k, j) as shown in Figure 6 (a) and (b). Moreover, we consider two hypothesis tests of pathways



(a) AD-MCI (b) CN

(c) AD-MCI (d) CN

Figure 6: Display of the subnetworks associated with genes APP and CASP3. (a) and (b): Solid/dashed arrows indicate significant/insignificant edges at $\alpha = 0.05$ after adjustment for multiplicity by the Bonferroni-Holm correction. (c) and (d): Solid arrows indicate the reconstructed edges using 2SPLS (Chen et al., 2018).

(a) AD-MCI (b) CN

Figure 7: The p-values of pathway tests (3) by the proposed tests for the AD-MCI and CN groups, where p-values are adjusted for multiplicity by the Bonferroni-Holm correction and solid/dashed arrows indicate significant/insignificant pathways at $\alpha = 0.05$.

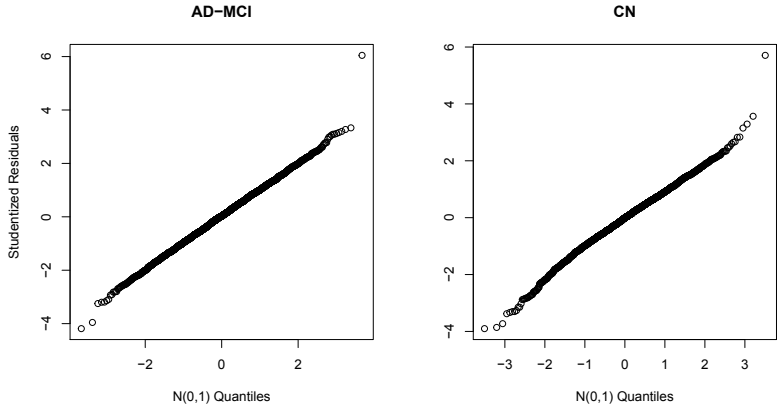


Figure 8: Normal quantile-quantile plots of studentized residuals of the AD-MCI and CN groups.

$H_0 : \mathrm{U}_{kj} = 0$ for some $(k, j) \in \mathcal{P}_\ell$ versus $H_a : \mathrm{U}_{kj} \neq 0$ for all $(k, j) \in \mathcal{P}_\ell$; $\ell = 1, 2$, where the two pathways are specified by $\mathcal{P}_1 = \{\text{PSEN1} \to \text{CAPN1} \to \text{CDK5R1}\}$, and $\mathcal{P}_2 = \{\text{PSEN1} \to \text{CAPN2} \to \text{CDK5R1}\}$. See Figure 7. Of note, for clear visualization, Figure 6 (a)-(b), and Figure 7 only display the edges related to hypothesis testing. Also notice that the ancestral relations are reconstructed using $p = 63$ genes and $q = 126$ SNPs for AD-MCI and CN groups separately.

In Figures 6-7, the significant results under the level $\alpha = 0.05$ after the Holm-Bonferroni adjustment for $2 \times (9 + 2) = 22$ tests are displayed. In Figures 6, the edge test in (2) exhibits a strong evidence for the presence of directed connectivity {APP $\to$ APBB1, APP $\to$ GSK3B, FADD $\to$ CASP3} in the AD-MCI group, but no evidence in the CN group. Meanwhile, this test suggests the presence of connections {TNFRSF1A $\to$ CASP3, FADD $\to$ CASP8} in the CN group but not so in the AD-MCI group. In both groups, we identify directed connections {TNFRSF1A $\to$ FADD, TNFRSF1A $\to$ CASP8}. In Figure 7, the pathway test (3) supports the presence of a pathway PSEN1 $\to$ CAPN1 $\to$ CDK5R1 in the AD-MCI group with a p-value of 0.044 but not in the CN group with a p-value of 0.33. The pathway PSEN1 $\to$ CAPN2 $\to$ CDK5R1 appears insignificant at $\alpha = 0.05$ for both groups. Also noted is that some of our discoveries agree with the literature according to the AlzGene database (`alzgene.org`) and the AlzNet database (`https://mips.helmholtz-muenchen.de/AlzNet-DB`). Specifically, GSK3B differentiates AD patients from normal subjects; as shown in Figures 6, our result indicates the presence of connection APP $\to$ GSK3B for the AD-MCI group, but not for the CN group, the former of which is confirmed by Figure 1 of Kremer et al. (2011). The connection APP $\to$ APBB1 also differs in AD-MCI and CN groups, which appears consistent with Figure 3 of Bu (2009). Moreover, the connection CAPN1 $\to$ CDK5R1, in the pathway PSEN1 $\to$ CAPN1 $\to$ CDK5R1 discovered in AD-MCI group, is found in the AlzNet database (interaction-ID 24614, `https://mips.helmholtz-muenchen.de/AlzNet-DB/entry/show/1870`). Finally, as suggested by Figure 8, the normality assumption in (1) is adequate for both groups.

By comparison, as shown in Figure 6 (c) and (d), gene APP in the reconstructed networks by 2SPLS (Chen et al., 2018) is not connected with other genes, indicating no regulatory relation of APP with other genes in the AD-MCI and CN groups. However, as a well-known gene associated with AD, APP is reported to play different roles in controlling the expressions of other genes for AD patients and healthy people (Matsui et al., 2007; Julia and Goate, 2017). Our results in Figure 6 (a) and (b) are congruous with the studies: the connections of APP with other genes are different in our estimated networks for AD-MCI and CN groups.

In summary, our findings seem to agree with those in the literature (Julia and Goate, 2017; Su et al., 2001; Kelleher III and Shen, 2017), where the subnetworks of genes APP, CASP3 in Figure 6 and PSEN1 in Figure 7 differentiate the AD-MCI from the CN groups. Furthermore, the pathway PSEN1 $\rightarrow$ CAPN1 $\rightarrow$ CDK5R1 in Figure 7 seems to differentiate these groups, which, however, requires validation in biological experiments.

7. Summary

This article proposes structure learning and inference methods for a Gaussian DAG with interventions, where the targets and strengths of interventions are unknown. A likelihood ratio test is derived based on an estimated ARG formed by ancestral relations and candidate interventional relations. This test accounts for the statistical uncertainty of the construction of the ARG based on a novel data perturbation scheme. Moreover, we develop a peeling algorithm for the ARG construction. The peeling algorithm allows scalable computing and yields a consistent estimator. The numerical studies justify our theory and demonstrate the utility of our methods.

The proposed methods can be extended to many practical situations beyond biological applications with independent and identically distributed data. An instance is to infer directed relations between multiple autoregressive time series (Pamfil et al., 2020), where the lagged variables and covariates can serve as interventions for each time series.

The current work has two limitations. First, the inferential theory requires (asymptotically) correct recovery of the local DAG structures (Remark 17) to produce valid p-values, similar to Shi et al. (2019) and Zhu et al. (2020). As illustrated in numerical studies, the graph structures are reasonably recovered when n is moderately large, and the DP scheme empirically alleviates the issue of inference after the ARG reconstruction. However, whether valid p-values can be obtained without the exact reconstruction of nuisance graph structures remains unclear in theory. Second, the proposed methods do not treat hidden confounding, which often arises in practice and can bias the results of both inference and learning. One future research direction is to extend the framework of unspecified interventions to allow unmeasured confounders.

Acknowledgments

The authors would like to thank the action editor and three anonymous reviewers for their helpful comments and suggestions. The research is supported by NSF grants DMS-

1712564, DMS-1952539, and NIH grants R01GM126002, R01HL116720, R01HL105397, R01AG069895, R01AG065636, R01AG074858, and U01AG073079.

Appendix A. Illustrative Examples and Discussions

A.1 Identifiability of Model (1) and Assumption 1

The parameter space for model (1) is

$$\{(\mathbf{U}, \mathbf{W}, \mathbf{\Sigma}) : \mathbf{U} \in \mathbb{R}^{p\times p} \text{ represents a DAG}, \ \mathbf{W} \in \mathbb{R}^{p\times q}, \ \mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)\}.$$

As suggested by Proposition 3, Assumption 1 (1A-1C) suffices for identification of every parameter value in the parameter space. Next, we show by examples that if Assumption 1B or 1C is violated then model (1) is no longer identifiable. To proceed, we rewrite (1) as

$$\boldsymbol{Y} = \mathbf{V}^\top \boldsymbol{X} + \boldsymbol{\varepsilon}_V, \quad \boldsymbol{\varepsilon}_V = (\mathbf{I} - \mathbf{U}^\top)^{-1}\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{\Omega}^{-1}),$$

where $\mathbf{\Omega} = (\mathbf{I} - \mathbf{U})\mathbf{\Sigma}^{-1}(\mathbf{I} - \mathbf{U}^\top)$ is a precision matrix and $\mathbf{V} = \mathbf{W}(\mathbf{I} - \mathbf{U})^{-1}$.

Example 2 (Identifiability) *In model* (1)*, consider two non-identifiable bivariate situations: (1)* $p = 2$ *and* $q = 4$ *and (2)* $p = 2$ *and* $q = 2$*.*

(1) Model (1) *is non-identifiable when Assumption 1B breaks down. Consider two different models with different parameter values:*

$$\boldsymbol{\theta}: \quad Y_1 = X_1 + X_2 + X_3 + \varepsilon_1, \qquad Y_2 = Y_1 - X_2 + X_3 + X_4 + \varepsilon_2 \tag{19}$$

$$\widetilde{\boldsymbol{\theta}}: \quad Y_1 = 0.5Y_2 + 0.5X_1 + X_2 - 0.5X_4 + \widetilde{\varepsilon}_1, \qquad Y_2 = X_1 + 2X_3 + X_4 + \widetilde{\varepsilon}_2, \tag{20}$$

where $\varepsilon_1, \varepsilon_2 \sim N(0, 1)$ *are independent, and* $\widetilde{\varepsilon}_1 \sim N(0, 0.5)$*,* $\widetilde{\varepsilon}_2 \sim N(0, 2)$ *are independent. As depicted in Figure 9,* (19) *satisfies Assumption 1C. However, Assumption 1B is violated given that* $\text{Cov}(Y_2, X_2 \mid \boldsymbol{X}_{\{1,3,4\}}) = 0$ *and* X_2 *is an intervention variable of* Y_2*. This is because the direct interventional effect of* X_2 *on* Y_2 *are canceled out by its indirect interventional effect through* Y_1*. Similarly,* (20) *satisfies Assumption 1C but* $\text{Cov}(Y_1, X_4 \mid \boldsymbol{X}_{\{1,2,3\}}) = 0$ *violating Assumption 1B. In this case, it can be verified that* $\boldsymbol{\theta}$ *and* $\widetilde{\boldsymbol{\theta}}$ *correspond to the same distribution* $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$*, because they share the same* $(\mathbf{V}, \mathbf{\Omega})$ *even with different values of* $(\mathbf{U}, \mathbf{W}, \mathbf{\Sigma})$*. Hence, it is impossible to infer the directed relation between* Y_1 *and* Y_2*.*

(2) Model (1) *is non-identifiable when Assumption 1C breaks down. Consider two different models with different parameter values:*

$$\boldsymbol{\theta}: \quad Y_1 = X_1 + X_2 + \varepsilon_1, \qquad Y_2 = Y_1 + X_2 + \varepsilon_2, \tag{21}$$

$$\widetilde{\boldsymbol{\theta}}: \quad Y_1 = 0.5Y_2 + 0.5X_1 + \widetilde{\varepsilon}_1, \qquad Y_2 = X_1 + 2X_2 + \widetilde{\varepsilon}_2, \tag{22}$$

where $\varepsilon_1, \varepsilon_2 \sim N(0, 1)$ *are independent, and* $\widetilde{\varepsilon}_1 \sim N(0, 0.5)$*,* $\widetilde{\varepsilon}_2 \sim N(0, 2)$ *are independent. Note that* (21) *and* (22) *satisfy Assumption 1B. In* (21)*,* Y_2 *does not have any instrumental intervention although it has an invalid instrument* X_2*. Similarly, in* (22)*, neither does* Y_1 *have any instrumental intervention while having an invalid instrument* X_1*. As in the previous case,* $\boldsymbol{\theta}$ *and* $\widetilde{\boldsymbol{\theta}}$ *yield the same distribution* $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$ *because they share the same* $(\mathbf{V}, \mathbf{\Omega})$ *even with different values of* $(\mathbf{U}, \mathbf{W}, \mathbf{\Sigma})$*. In this case, it is impossible to infer the directed relation between* Y_1 *and* Y_2*.*

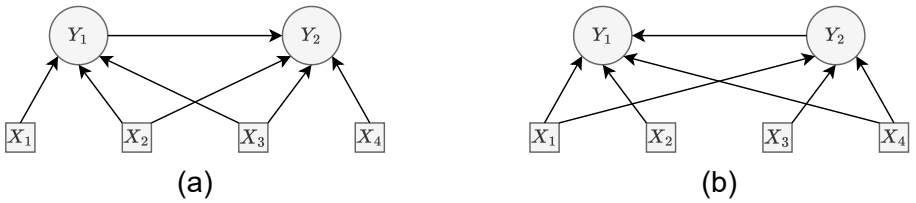


Figure 9: (a) Display of DAG defined by (19). (b) Display of DAG defined by (20).

Figure 10: (a) Display of DAG defined by (21). (b) Display of DAG defined by (22).

A.2 Illustration of Algorithm 2

We now illustrate Algorithm 2 by Example 3.

Example 3 *Consider model* (1) *with* $p = q = 5$,

$$\begin{aligned} &Y_1 = X_1 + \varepsilon_1, && Y_2 = 0.5Y_1 + X_3 + \varepsilon_2, && Y_5 = X_4 + \varepsilon_5, \\ &Y_3 = 0.5Y_2 + X_5 + \varepsilon_3, && Y_4 = 0.5Y_3 - 0.1Y_1 + X_2 + \varepsilon_4, \end{aligned} \tag{23}$$

where $\varepsilon_1, \ldots, \varepsilon_5 \sim N(0,1)$ *independently. Then* (23) *defines a DAG as displayed in Figure 1. For illustration, we generate a random sample of size* $n = 40$ *and compute* $\widehat{\mathbf{V}}$ *by Algorithm 1. In particular,*

$$\mathbf{V} = \begin{pmatrix} 1 & 0.5 & 0.25 & 0.025 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0.5 & 0 \end{pmatrix}, \quad \widehat{\mathbf{V}} = \begin{pmatrix} 0.92 & 0.48 & 0.27 & 0 & 0 \\ 0 & 0 & 0 & 1.08 & 0 \\ 0 & 1.03 & 0.52 & 0.21 & 0 \\ 0 & 0 & 0 & 0 & 1.06 \\ 0 & 0 & 0.98 & 0.55 & 0 \end{pmatrix}.$$

Algorithm 2 proceeds as follows.

- $\mathcal{V}_Y = \{1,2,3,4,5\}$*: The interventions indexed by* $\mathcal{B} = \{2,4\}$ *are instruments on the leaves that are indexed by* $\mathcal{L} = \{4,5\}$.
  - X_2 *is identified as an instrument of leaf node* Y_4 *(*$X_2 \to Y_4$*) because* $\widehat{\mathrm{V}}_{24} \neq 0$ *is the only nonzero in the row 2 with the smallest (positive) row* ℓ_0*-norm.*
  - X_4 *is identified as an instrument of leaf node* Y_5 *(*$X_4 \to Y_5$*) because* $\widehat{\mathrm{V}}_{45} \neq 0$ *is the only nonzero in row 4 with the smallest (positive) row* ℓ_0*-norm.*

  Then $\{Y_4, Y_5\}$ *are removed.*

- $\mathcal{V}_Y = \{1,2,3\}$*: The intervention indexed by* $\mathcal{B} = \{5\}$ *is an instrument on the leaf indexed by* $\mathcal{L} = \{3\}$*.*

  – X_5 *is identified as an instrument of a leaf node* Y_3 *(*$X_5 \to Y_3$*) in* $\mathcal{G}^{\text{work}}$ *given that* $\widehat{\mathrm{V}}_{53} \neq 0$ *is the only nonzero element in the row with the smallest (positive) row* ℓ_0*-norm of the submatrix for* Y_1, Y_2, Y_3*.*

  Since Y_3 *is a leaf in* $\mathcal{G}^{\text{work}}$*,* Y_4 *has been removed,* X_5 *is the only instrument on* Y_3*, and* $\widehat{\mathrm{V}}_{54} \neq 0$*, we have* $(3,4) \in \widehat{\mathcal{E}}_+$ *by Proposition 7. Then* $\{Y_3\}$ *is removed.*

- $\mathcal{V}_Y = \{1,2\}$*: The intervention indexed by* $\mathcal{B} = \{3\}$ *is an instrument on the leaf indexed by* $\mathcal{L} = \{2\}$*.*

  – X_3 *is identified as an instrument of a leaf node* Y_2 *(*$X_3 \to Y_2$*) similarly in* $\mathcal{G}^{\text{work}}$ *given that* $\widehat{\mathrm{V}}_{32} \neq 0$ *is the largest nonzero element in its row of the submatrix.*

  Since Y_2 *is a leaf in* $\mathcal{G}^{\text{work}}$*,* Y_3 *has been removed,* X_3 *is the only instrument on* Y_2*, and* $\widehat{\mathrm{V}}_{33} \neq 0$*, we have* $(2,3) \in \widehat{\mathcal{E}}_+$*. Then* $\{Y_2\}$ *is removed.*

- $\mathcal{V}_Y = \{1\}$*: The intervention indexed by* $\mathcal{B} = \{1\}$ *is an instrument on the leaf indexed by* $\mathcal{L} = \{1\}$*.*

  – X_1 *is an instrument of* Y_1 *(*$X_1 \to Y_1$*).*

  Since Y_1 *is a leaf node in* $\mathcal{G}^{\text{work}}$*,* Y_2 *has been removed,* X_1 *is the only instrument on* Y_1*, and* $\widehat{\mathrm{V}}_{12} \neq 0$*, we have* $(1,2) \in \widehat{\mathcal{E}}_+$*. Then* $\{Y_1\}$ *is removed, and the peeling process is terminated.*

Finally, Steps 9 and 10 identify

$$\widehat{\mathcal{E}}_+ = \{(1,2),(2,3),(3,4),(1,3),(1,4),(2,4)\},$$
$$\widehat{\mathcal{I}}_+ = \{(1,1),(1,2),(1,3),(1,4),(2,4),(3,2),(3,3),(3,4),(4,5),(5,3),(5,4)\},$$

which are equal to $\mathcal{E}_+$ *and* $\mathcal{I}_+$*, respectively.*

In Example 3, $\{(l,j) : \widehat{\mathrm{V}}_{lj} \neq 0\} \neq \{(l,j) : \mathrm{V}_{lj} \neq 0\}$, suggesting that the selection consistency of $\widehat{\mathbf{V}}$ is unnecessary for Algorithm 2 to correctly reconstruct the ARG $\mathcal{G}_+$; see also Section A.3 for the theoretical justification.

A.3 Relaxation of Assumption 3

Assumption 3 in Theorem 14 leads to consistent identification for $\mathbf{V}$. Now, we discuss when Algorithm 2 correctly reconstructs $\mathcal{G}_+$ without requiring Assumption 3.

Assumption 5 *For* $1 \le j \le p$*, there exists* τ_j^* *such that*

(A) $\Big\{l : X_l \text{ intervenes on } Y_j \text{ or its unmediated parents}\Big\} \subseteq \Big\{l : |\mathrm{V}_{lj}| \ge \tau_j^*\Big\}$.

(B) $\tau_j^* \ge 100 c_1^{-1} c_2\big(\Omega_{jj}^{-1/2}\big)\sqrt{(\kappa_j^\circ - |\{l : |\mathrm{V}_{lj}| \ge \tau_j^*\}| + 1)(\log(q)/n + \log(n)/n)}$.

Assumption 5 requires the effects of intervention variables on Y_j or its unmediated parents to exceed a certain signal strength τ_j^*, while imposing no restrictions on the other intervention variables, for $1 \le j \le p$. These signals enable us to reconstruct $\mathcal{G}_+$. Assumption 3 implies Assumption 5 with $\tau_j^* = \min_{\mathrm{V}_{lj} \neq 0} |\mathrm{V}_{lj}|$, so Assumption 5 is weaker.

Theorem 20 *Suppose Assumptions 1-2 and 5 are met with constants $c_1 < 6c_2$, and the machine precision* tol $\ll 1/n$ *is negligible. For $1 \le j \le p$, there exist some suitable choice of tuning parameters (κ_j, τ_j) in Algorithm 1 such that*

$$|\{l : |\mathrm{V}_{lj}| \ge \tau_j^*\}| \le \kappa_j \le \kappa_j^\circ, \qquad \frac{36c_2}{c_1}\sqrt{\Omega_{jj}^{-1}\Big(\frac{\log(q)}{n} + \frac{\log(n)}{n}\Big)} \le \tau_j \le \frac{2\tau_j^*}{5},$$

then for any γ_j such that

$$\tau_j^{-1}(32c_2^2\Omega_{jj}^{-1}n^{-1}(\log(q) + \log(n)))^{1/2} \le \gamma_j \le c_1/6\sqrt{\kappa_j^\circ - |\{l : |\mathrm{V}_{lj}| \ge \tau_j^*\}| + 1},$$

almost surely we have Algorithm 1 terminates in at most $1 + \lceil \log(\kappa_{\max}^\circ)/\log(4) \rceil$ DC iterations when n is sufficiently large. Moreover, almost surely we have Algorithm 2 recovers $\mathcal{E}_+$ and $\mathcal{I}_+$ when n is sufficiently large.

The proof of Theorem 20 is given in Appendix B.9.

A.4 Comparison of Strong Faithfulness and Assumption 3 (or 5)

In the literature, a faithfulness condition is usually assumed for identifiability up to Markov equivalence classes (Spirtes et al., 2000). For discussion, we formally introduce the concepts of faithfulness and strong faithfulness.

Consider a DAG $\mathcal{G}$ with node variables $(Z_1, \ldots, Z_{p+q})^\top$. Nodes Z_i and Z_j are adjacent if $Z_i \to Z_j$ or $Z_j \to Z_i$. A *path* (*undirected*) between Z_i and Z_j in $\mathcal{G}$ is a sequence of distinct nodes $(Z_i, \ldots, Z_j)$ such that all pairs of successive nodes in the sequence are adjacent. A nonendpoint node Z_k on a path $(Z_i, \ldots, Z_{k-1}, Z_k, Z_{k+1}, \ldots, Z_j)$ is a *collider* if $Z_{k-1} \to Z_k \leftarrow Z_{k+1}$. Otherwise it is a *noncollider*. Let $A \subseteq \{1, \ldots, p+q\}$, where A does not contain i and j. Then $\boldsymbol{Z}_A$ *blocks* a path $(Z_i, \ldots, Z_j)$ if at least one of the following holds: (i) the path contains a noncollider that is in $\boldsymbol{Z}_A$, or (ii) the path contains a collider that is not in $\boldsymbol{Z}_A$ and has no descendant in $\boldsymbol{Z}_A$. A node Z_i is d-separated from Z_j given $\boldsymbol{Z}_A$ if $\boldsymbol{Z}_A$ block every path between Z_i and Z_j; $i \neq j$ (Pearl, 2009).

According to Uhler et al. (2013), a multivariate Gaussian distribution of $(Z_1, \ldots, Z_{p+q})^\top$ is said to be ς-strong faithful to a DAG with node set $\mathcal{V} = \{1, \ldots, p+q\}$ if

$$\min_{A \subseteq \mathcal{V} \setminus \{i,j\}} \Big\{ |\mathrm{Corr}(Z_i, Z_j \mid \boldsymbol{Z}_A)| : Z_i \text{ is not d-separated from } Z_j \text{ given } \boldsymbol{Z}_A \Big\} > \varsigma, \tag{24}$$

for $1 \le i \ne j \le p+q$, where $\varsigma \in [0, 1)$, Corr denotes the correlation. When $\varsigma = 0$, (24) is equivalent to faithfulness. For consistent structure learning (up to Markov equivalence classes), it often requires that $\varsigma \gtrsim \sqrt{s_0 \log(p+q)/n}$, where s_0 is a sparsity measure; see Uhler et al. (2013) for a survey. For a pair (i, j), the number of possible sets for A is $2^{(p+q-2)}$. If $Z_i \to Z_j$, then $\mathrm{Corr}(Z_i, Z_j \mid \boldsymbol{Z}_A) \neq 0$ for any A. Therefore, for this (i, j) pair

alone, (24) could require exponentially many conditions. Indeed, (24) is very restrictive in high-dimensional situation (Uhler et al., 2013).

By comparison, Algorithm 2 yields consistent structure learning based on Assumption 3 or 5 instead of strong faithfulness. In some sense, Assumption 3 or 5 requires sufficient signal strength that is analogous to the condition for consistent feature selection (Shen et al., 2012). This assumption may be thought of as an alternative to strong faithfulness. As illustrated in Example 4, Assumption 3 or 5 is less stringent than strong faithfulness.

Example 4 (Faithfulness) *Assume* $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I})$. *Consider model* (1) *with* $p = q = 3$,

$$Y_1 = \mathrm{W}_{11}X_1 + \varepsilon_1, \quad Y_2 = \mathrm{U}_{12}Y_1 + \mathrm{W}_{22}X_2 + \varepsilon_2, \quad Y_3 = \mathrm{U}_{13}Y_1 + \mathrm{U}_{23}Y_2 + \mathrm{W}_{33}X_3 + \varepsilon_3,$$

where $\varepsilon_1, \varepsilon_2, \varepsilon_3 \sim N(0,1)$ *are independent and* $\mathrm{U}_{12}, \mathrm{U}_{13}, \mathrm{U}_{23}, \mathrm{W}_{11}, \mathrm{W}_{22}, \mathrm{W}_{33} \neq 0$. *Denote* $\boldsymbol{Z} = (Y_1, Y_2, Y_3, X_1, X_2, X_3)^\top$. *Since the directed relations among* $\boldsymbol{X}$ *are not of interest,* (24) *becomes*

$$\begin{aligned} &\min_{\substack{\boldsymbol{Z}_A=(\boldsymbol{Y}_{A_1},\boldsymbol{X}_{A_2}):\\ A_1\subseteq\{i,j\}^c, A_2}} \Big\{ |\mathrm{Corr}(Y_i, Y_j \mid \boldsymbol{Z}_A)| : Y_i, Y_j \text{ are not d-separated given } \boldsymbol{Z}_A \Big\} > \varsigma, \\ &\min_{\substack{\boldsymbol{Z}_A=(\boldsymbol{Y}_{A_1},\boldsymbol{X}_{A_2}):\\ A_1\subseteq\{j\}^c, A_2\subseteq\{l\}^c}} \Big\{ |\mathrm{Corr}(Y_j, X_l \mid \boldsymbol{Z}_A)| : Y_j, X_l \text{ are not d-separated given } \boldsymbol{Z}_A \Big\} > \varsigma, \end{aligned} \tag{25}$$

for each pair (i, j) *with* $i \neq j$ *and each pair* (l, j). *Then strong-faithfulness in* (25) *assumes* 152 *conditions for the correlations. By comparison, Assumption 3 requires the absolute values of* $\mathrm{V}_{11}, \mathrm{V}_{12}, \mathrm{V}_{13}, \mathrm{V}_{22}, \mathrm{V}_{23}, \mathrm{V}_{33} \gtrsim \sqrt{\log(q)/n}$, *which in turn requires the minimum absolute value of six correlations* $\gtrsim \sqrt{\log(q)/n}$,
(i) $\mathrm{Corr}(Y_1, X_1 \mid X_2, X_3)$, *(ii)* $\mathrm{Corr}(Y_2, X_1 \mid X_2, X_3)$, *(iii)* $\mathrm{Corr}(Y_3, X_1 \mid X_2, X_3)$,
(iv) $\mathrm{Corr}(Y_2, X_2 \mid X_1, X_3)$, *(v)* $\mathrm{Corr}(Y_3, X_2 \mid X_1, X_3)$, *(vi)* $\mathrm{Corr}(Y_3, X_3 \mid X_1, X_2)$.
Importantly, (i)-(vi) are required in (25), *suggesting that the strong-faithfulness is more stringent than Assumption 3.*

A.5 Irregular Hypothesis

Assume, without loss of generality, that $\widehat{\mathcal{D}} = \mathcal{D}$ and $\widehat{\mathcal{G}}_+ = \mathcal{G}$ are correctly reconstructed in the following discussion.

- For testing of directed edges (2), suppose H_0 is irregular, namely, $\mathcal{D} \cup \mathcal{E}$ contains a directed cycle. This implies that a directed cycle exists in $\widehat{\mathcal{D}} \cup \widehat{\mathcal{E}}_+$. In this situation, we decompose H_0 into sub-hypotheses $H_0^{(1)}, \ldots, H_0^{(\nu)}$, each of which is regular. Then testing H_0 is equivalent to multiple testing for $H_0^{(1)}, \ldots, H_0^{(\nu)}$. For instance, in Example 1, $H_0 : \mathrm{U}_{45} = \mathrm{U}_{53} = 0$ is irregular, and H_0 can be decomposed into $H_0^{(1)} : \mathrm{U}_{45} = 0$ and $H_0^{(2)} : \mathrm{U}_{53} = 0$.

- For testing of directed pathways (3), if H_0 is irregular, then $\widehat{\mathcal{D}} \cup \widehat{\mathcal{E}}_+$ has a directed cycle. The p-value is defined to be one in this situation since no evidence supports the presence of the pathway.

A.6 Theoretical Results on Structure Learning

The regression (9) can be solved by Algorithm 1 with the input $\mathbf{Y}_{\cdot j}$ as the response variable and the input $(\mathbf{Y}_{\cdot,\text{AN}_{\widehat{\mathcal{G}}_+}(j)}, \mathbf{X}_{\cdot,\text{IN}_{\widehat{\mathcal{G}}_+}(j)})$ as the covariates for $1 \le j \le p$. The tuning parameters for solving (9) by Algorithm 1 are denoted by $\{(\kappa'_j, \tau'_j)\}_{1\le j\le p}$.

Let $\overline{\kappa} = \max_{1\le j\le p} |\text{AN}_{\mathcal{G}_+}(j)| + |\text{IN}_{\mathcal{G}_+}(j)|$ and $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$.

Assumption 6 *For constants $c_3, c_4 > 0$,*

(A) $\min_{\{A:|A|\le 2\overline{\kappa}\}} \min_{\{\boldsymbol{\zeta}:\|\boldsymbol{\zeta}_{A^c}\|_1 \le 3\|\boldsymbol{\zeta}_A\|_1\}} \|\mathbf{Z}\boldsymbol{\zeta}\|_2^2/n\|\boldsymbol{\zeta}\|_2^2 \ge c_3$ *almost surely.*

(B) $\max_{1\le k\le p+q} n^{-1}(\mathbf{Z}^\top\mathbf{Z})_{kk} \le c_4^2$ *almost surely.*

Assumption 7 $\min_{\text{U}_{kj}\neq 0} |\text{U}_{kj}| \ge 100c_3^{-1}c_4 \max_{1\le j\le p}(\sigma_j)\sqrt{\log(p)/n + \log(n)/n}$.

Theorem 21 *Suppose the assumptions in Theorem 14 are satisfied. In addition, suppose Assumptions 6-7 are met with constants $c_3 < 6c_4$. For $1 \le j \le p$, assuming $|\text{AN}_{\mathcal{G}}(j)| \ll n$ and $|\text{IN}_{\mathcal{G}}(j)| \ll n$, if the tuning parameters (κ'_j, τ'_j) are suitably chosen such that*

$$\kappa'_j = |\text{PA}_{\mathcal{G}}(j)|, \qquad \frac{36c_4}{c_3}\sigma_j\sqrt{\frac{\log(p)}{n} + \frac{\log(n)}{n}} \le \tau'_j \le \frac{2}{5}\min_{\text{U}_{kj}\neq 0}|\text{U}_{kj}|,$$

then for any γ_j such that $(\tau'_j)^{-1}(32c_4^2\sigma_j^2 n^{-1}(\log(p)+\log(n)))^{1/2} \le \gamma_j \le c_3/6$, almost surely we have $\widehat{\mathcal{E}} = \mathcal{E}$, when n is sufficiently large.

Appendix B. Technical Proofs

B.1 Proof of Proposition 3

Suppose that $\boldsymbol{\theta} = (\mathbf{U}, \mathbf{W}, \boldsymbol{\Sigma})$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{U}}, \tilde{\mathbf{W}}, \tilde{\boldsymbol{\Sigma}})$ render the same distribution of $(\boldsymbol{Y}, \boldsymbol{X})$. We will prove that $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

Denote by $\mathcal{G}(\boldsymbol{\theta})$ and $\mathcal{G}(\tilde{\boldsymbol{\theta}})$ the DAGs corresponding to $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$, respectively. First, consider $\mathcal{G}(\boldsymbol{\theta})$. Without loss of generality, assume Y_1 is a leaf node in $\mathcal{G}(\boldsymbol{\theta})$. By Assumption 1C, there exists an instrumental intervention with respect to $\mathcal{G}(\boldsymbol{\theta})$, say X_1. Then,

$$\text{Cov}(Y_j, X_1 \mid \boldsymbol{X}_{\{2,\dots,q\}}) = 0, \qquad j = 2,\dots,p, \tag{26}$$

$$\text{Cov}(Y_1, X_1 \mid \boldsymbol{Y}_A, \boldsymbol{X}_{\{2,\dots,q\}}) \neq 0, \qquad \text{for any } A \subseteq \{2,\dots,p\}. \tag{27}$$

By the local Markov property (Spirtes et al., 2000), (27) implies that $X_1 \to Y_1$ in $\mathcal{G}(\tilde{\boldsymbol{\theta}})$. Suppose Y_1 is not a leaf node in $\mathcal{G}(\tilde{\boldsymbol{\theta}})$. Without loss of generality, assume that Y_1 is an unmediated parent of Y_2. Then $\text{Cov}(Y_2, X_1 \mid \boldsymbol{X}_{\{2,\dots,q\}}) = 0$ but $X_1 \to Y_1$ and Y_1 is an unmediated parent of Y_2, which contradicts to Assumption 1B. This implies that if Y_1 is a leaf node in $\mathcal{G}(\boldsymbol{\theta})$ then it must be a leaf node in $\mathcal{G}(\tilde{\boldsymbol{\theta}})$. In both $G(\boldsymbol{\theta})$ and $\mathcal{G}(\tilde{\boldsymbol{\theta}})$, the parents and interventions of Y_1 can be identified by

$$\mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\{2,\dots,p\}}, \boldsymbol{X}) = \mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\text{PA}_{\mathcal{G}(\boldsymbol{\theta})}(1)}, \boldsymbol{X}) = \mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\text{PA}_{\mathcal{G}(\boldsymbol{\theta})}(1)}, \boldsymbol{X}_{\text{IN}_{\mathcal{G}(\boldsymbol{\theta})}(1)}),$$
$$\mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\{2,\dots,p\}}, \boldsymbol{X}) = \mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\text{PA}_{\mathcal{G}(\tilde{\boldsymbol{\theta}})}(1)}, \boldsymbol{X}) = \mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\text{PA}_{\mathcal{G}(\tilde{\boldsymbol{\theta}})}(1)}, \boldsymbol{X}_{\text{IN}_{\mathcal{G}(\tilde{\boldsymbol{\theta}})}(1)}).$$

Consequently, Y_1 has the same parents and interventions in $\mathcal{G}(\boldsymbol{\theta})$ and $\mathcal{G}(\tilde{\boldsymbol{\theta}})$.

The forgoing argument is applied to other nodes sequentially. First, we remove Y_1 with any directed edges to Y_1, which does not alter the joint distribution of $(\boldsymbol{Y}_{\{2,\ldots,p\}}, \boldsymbol{X})$ and the sub-DAG of nodes $Y_2, \ldots, Y_p$. By induction, we remove the leafs in $\mathcal{G}(\boldsymbol{\theta})$ until it is empty, leading to $\mathcal{G}(\boldsymbol{\theta}) = \mathcal{G}(\tilde{\boldsymbol{\theta}})$. Finally, $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ because they have the same locations for nonzero elements and these model parameters (or regression coefficients) are uniquely determined under Assumption 1A (Shojaie and Michailidis, 2010). This completes the proof.

B.2 Proof of Proposition 5

Proof of (A)

Note that the maximal length of a path in a DAG of p nodes is at most $p-1$. Then it can be verified that $\mathbf{U}$ is nilpotent in that $\mathbf{U}^p = \mathbf{0}$. An application of the matrix series expansion yields that $(\mathbf{I}-\mathbf{U})^{-1} = \mathbf{I}+\mathbf{U}+\cdots+\mathbf{U}^{p-1}$. Using the fact that $\mathbf{V} = \mathbf{W}(\mathbf{I}-\mathbf{U})^{-1}$ from (6), we have, for any $1 \le l, j \le p$,

$$\mathrm{V}_{lj} = \sum_{k=1}^{p} \mathrm{W}_{lk}\Big(\mathrm{I}_{kj} + \mathrm{U}_{kj} + \cdots + (\mathbf{U}^{p-1})_{kj}\Big),$$

where U_{kj} is the (k,j)th entry of $\mathbf{U}$. If $\mathrm{V}_{lj} \neq 0$, then there exists k such that $W_{lk} \neq 0$ and $(\mathbf{U}^r)_{kj} \neq 0$ for some $0 \le r \le p-1$. If $r = 0$, then we must have $k = j$, and $X_l \to Y_j$. If $r > 0$, then $X_l \to Y_k$ and Y_k is an ancestor of Y_j.

Proof of (B)

First, for any leaf node variable Y_j, by Assumption 1, there exists an instrument $X_l \to Y_j$. If $\mathrm{V}_{lj'} \neq 0$ for some $j' \neq j$, then Y_j must be an ancestor of $Y_{j'}$, which contradicts the fact that Y_j is a leaf node variable.

Conversely, suppose that $\mathrm{V}_{lj} \neq 0$ and $\mathrm{V}_{lj'} = 0$ for $j' \neq j$. If Y_j is not a leaf node variable, then there exists a variable $Y_{j'}$ such that Y_j is an unmediated parent of $Y_{j'}$, that is $\mathrm{U}_{jj'} \neq 0$ and $(\mathbf{U}^r)_{jj'} = 0$ for $r > 1$. Then $\mathrm{V}_{lj'} = \mathrm{W}_{lj}\mathrm{U}_{jj'} \neq 0$, a contradiction.

B.3 Proof of Proposition 7

Suppose Y_k is an unmediated parent of Y_j. Let X_l be an instrument of Y_k in $\mathcal{G}_{\mathcal{L}^c}$. Then there are two cases: (1) X_l intervenes on Y_k but does not intervene on Y_j, namely $X_l \to Y_k$ but $X_l \not\to Y_j$; (2) X_l intervenes on Y_k and Y_j simultaneously, namely $X_l \to Y_k$ and $X_l \to Y_j$. For (1), $\mathrm{V}_{lj} = \mathrm{W}_{lk}\mathrm{U}_{kj} \neq 0$. For (2), Assumption 1B implies that $\mathrm{V}_{lj} \neq 0$. This holds for every instrument of Y_k in $\mathcal{G}_{\mathcal{L}^c}$, and the desired result follows.

Conversely, suppose for each instrument X_l of Y_k in $\mathcal{G}_{\mathcal{L}^c}$, we have $\mathrm{V}_{lj} \neq 0$. Let $X_{l'}$ be an instrument of Y_k in $\mathcal{G}$, which is also an instrument in $\mathcal{G}_{\mathcal{L}^c}$. Then $0 \neq \mathrm{V}_{l'j} = \mathrm{W}_{l'k}\mathrm{U}_{kj}$, which implies $\mathrm{U}_{kj} \neq 0$. This completes the proof.

B.4 Proof of Theorem 14

The proof proceeds in two steps: (A) we show that $\{l : \widehat{\mathrm{V}}_{lj} \neq 0\} = \{l : \mathrm{V}_{lj} \neq 0\}$ almost surely for $1 \le j \le p$; and (B) we show that $\widehat{\mathcal{G}}_+ = \mathcal{G}_+$ if $\widehat{\mathbf{V}}$ satisfies the property in (A).

Proof of (A)

Let $A_j^\circ = \Big\{l : \mathrm{V}_{lj} \neq 0\Big\}$ and $A_j^{[t]} = \Big\{l : |\widetilde{\mathrm{V}}_{lj}^{[t]}| \geq \tau_j\Big\}$ be the estimated support of penalized solution at the t-th iteration of Algorithm 1. For the penalized solution, define the false negative set $\mathrm{FN}_j^{[t]} = A_j^\circ \setminus A_j^{[t]}$ and the false positive set $\mathrm{FP}_j^{[t]} = A_j^{[t]} \setminus A_j^\circ$ for $t \geq 0$. Consider a "good" event

$$\mathscr{E}_j = \Big\{\|\mathbf{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty \leq 0.5\gamma_j\tau_j\Big\} \cap \Big\{\|\widehat{\mathbf{V}}_{\cdot j}^\circ - \mathbf{V}_{\cdot j}\|_\infty \leq 0.5\tau_j\Big\},$$

where $\widehat{\boldsymbol{\xi}}_j = \mathbf{Y}_j - \mathbf{X}\widehat{\mathbf{V}}_{\cdot j}^\circ$ is the residual of the oracle least squares estimate $\widehat{\mathbf{V}}_{\cdot j}^\circ$ such that $A_j = \Big\{l : \widehat{\mathrm{V}}_{lj}^\circ \neq 0\Big\}$; $1 \leq j \leq p$. We shall show that $\mathrm{FN}_j^{[t]}$ and $\mathrm{FP}_j^{[t]}$ are eventually empty on event $\mathscr{E}_j$ which has a probability tending to one.

First, we will show that if $|A_j^\circ \cup A_j^{[t-1]}| \leq 2\kappa_{\max}^\circ$ on $\mathscr{E}_j$ for $t \geq 1$, then $|A_j^\circ \cup A_j^{[t]}| \leq 2\kappa_{\max}^\circ$, to be used in Assumption 2A. To proceed, suppose $|A_j^\circ \cup A_j^{[t-1]}| \leq 2\kappa_{\max}^\circ$ on $\mathscr{E}_j$ for $t \geq 1$. By the optimality condition of (8),

$$\Big(\widehat{\mathbf{V}}_{\cdot j}^\circ - \widetilde{\mathbf{V}}_{\cdot j}^{[t]}\Big)^\top \Big(- \mathbf{X}^\top(\mathbf{Y}_j - \mathbf{X}\widetilde{\mathbf{V}}_{\cdot j}^{[t]})/n + \gamma_j\tau_j\nabla\|\widetilde{\mathbf{V}}_{(A_j^{[t-1]})^c,j}^{[t]}\|_1\Big) \geq 0,$$

where $\widetilde{\mathbf{V}}^{[t]}$ is defined in (8). Plugging $\widehat{\boldsymbol{\xi}}_j = \mathbf{Y}_j - \mathbf{X}\widehat{\mathbf{V}}_{\cdot j}^\circ$ into the inequality and rearranging it, we have $\|\mathbf{X}(\widetilde{\mathbf{V}}_{\cdot j}^{[t]} - \widehat{\mathbf{V}}_{\cdot j}^\circ)\|_2^2/n$ is no greater than

$$\begin{aligned}
&\Big(\widetilde{\mathbf{V}}_{\cdot j}^{[t]} - \widehat{\mathbf{V}}_{\cdot j}^\circ\Big)^\top \Big(\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n - \gamma_j\tau_j\nabla\|\widetilde{\mathbf{V}}_{(A_j^{[t-1]})^c,j}^{[t]}\|_1\Big) \\
= &\Big(\widetilde{\mathbf{V}}_{A_j^\circ\setminus A_j^{[t-1]},j}^{[t]} - \widehat{\mathbf{V}}_{A_j^\circ\setminus A_j^{[t-1]},j}^\circ\Big)^\top \Big(\mathbf{X}_{A_j^\circ\setminus A_j^{[t-1]}}^\top\widehat{\boldsymbol{\xi}}_j/n - \gamma_j\tau_j\nabla\|\widetilde{\mathbf{V}}_{A_j^\circ\setminus A_j^{[t-1]},j}^{[t]}\|_1\Big) \\
&+ \Big(\widetilde{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^{[t]} - \widehat{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^\circ\Big)^\top \Big(\mathbf{X}^\top\widehat{\boldsymbol{\xi}}/n - \gamma_j\tau_j\nabla\|\widetilde{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^{[t]}\|_1\Big) \\
&+ \Big(\widetilde{\mathbf{V}}_{A_j^{[t-1]}\setminus A_j^\circ,j}^{[t]} - \widehat{\mathbf{V}}_{A_j^{[t-1]}\setminus A_j^\circ,j}^\circ\Big)^\top \mathbf{X}_{A_j^{[t-1]}\setminus A_j^\circ}^\top\widehat{\boldsymbol{\xi}}/n,
\end{aligned} \tag{28}$$

where $\mathbf{X}_{A_j^\circ}^\top\widehat{\boldsymbol{\xi}}_j/n = \mathbf{0}$ has been used. Note that

$$\Big(\widetilde{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^{[t]} - \widehat{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^\circ\Big)^\top \nabla\Big\|\widetilde{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^{[t]}\Big\|_1 = \Big\|\widetilde{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^{[t]} - \widehat{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^\circ\Big\|_1.$$

Then (28) is no greater than

$$\begin{aligned}
&\Big\|\widetilde{\mathbf{V}}_{A_j^\circ\triangle A_j^{[t-1]},j}^{[t]} - \widehat{\mathbf{V}}_{A_j^\circ\triangle A_j^{[t-1]},j}^\circ\Big\|_1\Big(\|\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty + \gamma_j\tau_j\Big) \\
&+ \Big\|\widetilde{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^{[t]} - \widehat{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^\circ\Big\|_1\Big(\|\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty - \gamma_j\tau_j\Big),
\end{aligned} \tag{29}$$

where $\triangle$ denotes the symmetric difference of two sets. Note that $\|\mathbf{X}(\widetilde{\mathbf{V}}_{\cdot j}^{[t]} - \widehat{\mathbf{V}}_{\cdot j}^\circ)\|_2^2/n \geq 0$. Rearranging the inequality yields that

$$\begin{aligned}
&\Big(\gamma_j\tau_j - \|\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty\Big)\Big\|\widetilde{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^{[t]} - \widehat{\mathbf{V}}_{(A_j^\circ\cup A_j^{[t-1]})^c,j}^\circ\Big\|_1 \\
&\leq \Big(\|\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty + \gamma_j\tau_j\Big)\Big\|\widetilde{\mathbf{V}}_{A_j^\circ\triangle A_j^{[t-1]},j}^{[t]} - \widehat{\mathbf{V}}_{A_j^\circ\triangle A_j^{[t-1]},j}^\circ\Big\|_1.
\end{aligned}$$

On event $\mathscr{E}_j$, $\|\mathbf{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty \le \gamma_j\tau_j/2$, implying that

$$\begin{aligned}\Big\|\widetilde{\mathbf{V}}^{[t]}_{(A_j^\circ\cup A_j^{[t-1]})^c,j} - \widehat{\mathbf{V}}^\circ_{(A_j^\circ\cup A_j^{[t-1]})^c,j}\Big\|_1 &\le 3\Big\|\widetilde{\mathbf{V}}^{[t]}_{A_j^\circ\triangle A_j^{[t-1]},j} - \widehat{\mathbf{V}}^\circ_{A_j^\circ\triangle A_j^{[t-1]},j}\Big\|_1\\ &\le 3\Big\|\widetilde{\mathbf{V}}^{[t]}_{A_j^\circ\cup A_j^{[t-1]},j} - \widehat{\mathbf{V}}^\circ_{A_j^\circ\cup A_j^{[t-1]},j}\Big\|_1.\end{aligned}$$

Note that $|A_j^\circ \cup A_j^{[t-1]}| \le 2\kappa^\circ_{\max}$. By Assumption 2A and (29),

$$\begin{aligned}c_1\|\widetilde{\mathbf{V}}^{[t]}_{\cdot j} - \widehat{\mathbf{V}}^\circ_{\cdot j}\|_2^2 \le& \Big(\|\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty + \gamma_j\tau_j\Big)\Big\|\widetilde{\mathbf{V}}^{[t]}_{A_j^\circ\triangle A_j^{[t-1]},j} - \widehat{\mathbf{V}}^\circ_{A_j^\circ\triangle A_j^{[t-1]},j}\Big\|_1\\ &+ \Big(\|\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty - \gamma_j\tau_j\Big)\Big\|\widetilde{\mathbf{V}}^{[t]}_{(A_j^\circ\cup A_j^{[t-1]})^c,j} - \widehat{\mathbf{V}}^\circ_{(A_j^\circ\cup A_j^{[t-1]})^c,j}\Big\|_1 \qquad (30)\\ \le& \Big(\|\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty + \gamma_j\tau_j\Big)\Big\|\widetilde{\mathbf{V}}^{[t]}_{A_j^\circ\triangle A_j^{[t-1]},j} - \widehat{\mathbf{V}}^\circ_{A_j^\circ\triangle A_j^{[t-1]},j}\Big\|_1.\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\Big\|\widetilde{\mathbf{V}}^{[t]}_{A_j^\circ\triangle A_j^{[t-1]},j} - \widehat{\mathbf{V}}^\circ_{A_j^\circ\triangle A_j^{[t-1]},j}\Big\|_1 \le \sqrt{A_j^\circ\triangle A_j^{[t-1]}}\Big\|\widetilde{\mathbf{V}}^{[t]}_{\cdot j} - \widehat{\mathbf{V}}^\circ_{\cdot j}\Big\|_2.$$

Thus, $c_1\|\widetilde{\mathbf{V}}^{[t]}_{\cdot j} - \widehat{\mathbf{V}}^\circ_{\cdot j}\|_2 \le 1.5\gamma_j\tau_j\sqrt{2\kappa^\circ_{\max}}$, since $|A_j^\circ\triangle A_j^{[t-1]}| \le |A_j^\circ \cup A_j^{[t-1]}| \le 2\kappa^\circ_{\max}$ and $\|\mathbf{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty \le 0.5\gamma_j\tau_j$. By the condition of Theorem 14, $\|\widetilde{\mathbf{V}}^{[t]}_{\cdot j} - \widehat{\mathbf{V}}^\circ_{\cdot j}\|_2/\tau_j \le \sqrt{\kappa^\circ_{\max}}$ and $\gamma_j \le c_1/6$. On the other hand, for any $l \in \mathrm{FP}_j^{[t]} = A_j^{[t]} \setminus A_j^\circ$, we have $|\widetilde{\mathrm{V}}^{[t]}_{lj} - \widehat{\mathrm{V}}^\circ_{lj}| = |\widetilde{\mathrm{V}}^{[t]}_{lj}| > \tau_j$. Thus, $\sqrt{|\mathrm{FP}_j^{[t]}|} \le \|\widetilde{\mathbf{V}}^{[t]}_{\cdot j} - \widehat{\mathbf{V}}^\circ_{\cdot j}\|_2/\tau_j \le \sqrt{\kappa^\circ_{\max}}$, implying $|A_j^\circ \cup A_j^{[t]}| = |A_j^\circ| + |\mathrm{FP}_j^{[t]}| \le 2\kappa^\circ_{\max}$ on $\mathscr{E}_j$ for $t \ge 1$.

Next, we estimate the number of iterations required for termination. Note that the machine precision is negligible. The termination criterion is met when $A_j^{[t]} = A_j^{[t-1]}$ since the weighted Lasso problem (8) remains same at the tth and $(t-1)$th iterations. To show that $A_j^{[t]} = A_j^\circ$ eventually, we prove that $|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}| < 1$ eventually.

Now, suppose $|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}| \ge 1$. For any $l \in \mathrm{FN}_j^{[t]} \cup \mathrm{FP}_j^{[t]}$, by Assumption 3,

$$|\widetilde{\mathrm{V}}^{[t]}_{lj} - \widehat{\mathrm{V}}^\circ_{lj}| \ge |\widetilde{\mathrm{V}}^{[t]}_{lj} - \mathrm{V}_{lj}| - |\widehat{\mathrm{V}}^\circ_{lj} - \mathrm{V}_{lj}| \ge \tau_j - 0.5\tau_j,$$

so $\sqrt{|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}|} \le \|\widetilde{\mathbf{V}}^{[t]}_{\cdot j} - \widehat{\mathbf{V}}^\circ_{\cdot j}\|_2/0.5\tau_j$. By (30) and the Cauchy-Schwarz inequality, $c_1\|\widetilde{\mathbf{V}}^{[t]}_{\cdot j} - \widehat{\mathbf{V}}^\circ_{\cdot j}\|_2 \le 1.5\gamma_j\tau_j\sqrt{|\mathrm{FN}_j^{[t-1]}| + |\mathrm{FP}_j^{[t-1]}|}$. By conditions (1) and (2) for (τ_j, γ_j) in Theorem 14, we have

$$\begin{aligned}\sqrt{|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}|} \le \frac{\|\widetilde{\mathbf{V}}^{[t]}_{\cdot j} - \widehat{\mathbf{V}}^\circ_{\cdot j}\|_2}{0.5\tau_j} &\le \frac{3\gamma_j}{c_1}\sqrt{|\mathrm{FN}_j^{[t-1]}| + |\mathrm{FP}_j^{[t-1]}|}\\ &\le 0.5\sqrt{|\mathrm{FN}_j^{[t-1]}| + |\mathrm{FP}_j^{[t-1]}|}.\end{aligned}$$

Hence, $\sqrt{|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}|} \le (1/2)^t\sqrt{|A_j^\circ| + |A_j^{[0]}|}$. In particular, for $t \ge 1 + \lceil \log \kappa_j^\circ / \log 4 \rceil$, $|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}| < 1$ implying that $\mathrm{FN}_j^{[t]} = \emptyset$ and $\mathrm{FP}_j^{[t]} = \emptyset$.

Let $t_{\max} = 1 + \lceil \log \kappa_j^\circ / \log 4 \rceil$. Then $\mathrm{FN}_j^{[t_{\max}]} = \mathrm{FP}_j^{[t_{\max}]} = \emptyset$ on event $\mathscr{E}_j$. By the condition of Theorem 14, we have $\Big\{ l : \widetilde{\mathrm{V}}_{lj}^{[t_{\max}]} \neq 0 \Big\} = \Big\{ l : \mathrm{V}_{lj} \neq 0 \Big\}$. To bound $\mathbb{P}\Big(\bigcup_{j=1}^p \mathscr{E}_j^c \Big)$, let $\boldsymbol{\eta} = \mathbf{X}^\top (\mathbf{I} - \mathbf{P}_{A_j^\circ}) \boldsymbol{\xi}_j$ and $\boldsymbol{\eta}' = n(\mathbf{X}_{A_j^\circ}^\top \mathbf{X}_{A_j^\circ})^{-1} \mathbf{X}_{A_j^\circ}^\top \boldsymbol{\xi}_j$, where $\boldsymbol{\xi}_j = ((\varepsilon_V)_{1j}, \ldots, (\varepsilon_V)_{nj})^\top$. Then $\boldsymbol{\eta} \in \mathbb{R}^q$ and $\boldsymbol{\eta}' \in \mathbb{R}^{|\kappa_j^\circ|}$ are Gaussian vectors with $\mathrm{Var}(\eta_l) \leq \Omega_{jj}^{-1} (\mathbf{X}^\top \mathbf{X})_{ll} \leq n \Omega_{jj}^{-1} c_2^2$ and $\mathrm{Var}(\eta_l') \leq \Omega_{jj}^{-1} ((\mathbf{X}_{A_j^\circ}^\top \mathbf{X}_{A_j^\circ})^{-1})_{ll} \leq n \Omega_{jj}^{-1} c_2^2$. Then

$$\begin{aligned}
\mathbb{P}\Big(\|\mathbf{X}^\top \widehat{\boldsymbol{\xi}}_j / n\|_\infty > 0.5 \gamma_j \tau_j \Big) &= \mathbb{P}\Big(\|\boldsymbol{\eta}/n\|_\infty > 0.5 \gamma_j \tau_j \Big) \\
&\leq \sum_{l=1}^q \mathbb{P}\Big(|\eta_l| \geq 0.5 n \gamma_j \tau_j \Big) \\
&\leq q \int_{\frac{0.5\sqrt{n\Omega_{jj}} \gamma_j \tau_j}{c_2}}^\infty \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt \leq q \sqrt{\frac{2}{\pi}} \exp\Big(- \frac{n \Omega_{jj} \gamma_j^2 \tau_j^2}{8 c_2^2} \Big),
\end{aligned}$$

and similarly,

$$\mathbb{P}\Big(\|\widehat{\mathbf{V}}_{\cdot j}^\circ - \mathbf{V}_{\cdot j}^\circ\|_\infty > 0.5 \tau_j \Big) \leq \sum_{l \in A_j^\circ} \mathbb{P}\Big(|\eta_l'| > 0.5 \tau_j \Big) \leq \kappa_{\max}^\circ \sqrt{\frac{2}{\pi}} \exp\Big(- \frac{n \Omega_{jj} \tau_j^2}{8 c_2^2} \Big).$$

Under conditions for (τ_j, γ_j) in Theorem 14,

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{ l : \widetilde{\mathrm{V}}_{lj}^{[t_{\max}]} \neq 0 \Big\} = \Big\{ l : \mathrm{V}_{lj} \neq 0 \Big\};\ 1 \leq j \leq p \Big) \\
&\geq 1 - \mathbb{P}\Big(\bigcup_{j=1}^p \mathscr{E}_j^c \Big) \\
&\geq 1 - p \sqrt{\frac{2}{\pi}} \Big(e^{-4(\log(q) + \log(n)) + \log(q)} + e^{-144 c_1^{-2} c_2^2 (\log(q) + \log(n)) + \log(\kappa_j^\circ)} \Big) \geq 1 - \sqrt{\frac{2}{\pi}} p q^{-3} n^{-4},
\end{aligned}$$

where $c_1 < 6 c_2$ is used in the last inequality. By Borel-Cantelli lemma, almost surely we have $\Big\{ l : \widetilde{\mathrm{V}}_{lj}^{[t_{\max}]} \neq 0 \Big\} = \Big\{ l : \mathrm{V}_{lj} \neq 0 \Big\}$; $1 \leq j \leq p$ when n is sufficiently large.

So far, we have $\widetilde{\mathbf{V}}_{\cdot j} = \widetilde{\mathbf{V}}_{\cdot j}^{[t_{\max}]} = \widehat{\mathbf{V}}_{\cdot j}^\circ$; $1 \leq j \leq p$, almost surely. It remains to show that $\widehat{\mathbf{V}}_{\cdot j}^\circ$ is a global minimizer of (7); $1 \leq j \leq p$, with a high probability. Note that Assumptions 2A and 3 imply the degree of separation condition (Shen et al., 2013). By Theorem 2 of Shen et al. (2013),

$$\mathbb{P}\Big(\widehat{\mathbf{V}}_{\cdot j}^\circ \text{ is not a global minimizer of (7)}; 1 \leq j \leq p \Big) \leq 3p \exp\Big(- 2(\log(q) + \log(n)) \Big).$$

implying that almost surely $\widehat{\mathbf{V}}_{\cdot j} = \widetilde{\mathbf{V}}_{\cdot j}^{[t_{\max}]} = \widehat{\mathbf{V}}_{\cdot j}^\circ$ is a global minimizer of (7) when n is sufficiently large. This completes the proof.

Proof of (B)

Assume $\widehat{\mathbf{V}}$ satisfies the properties in (A). Then Propositions 5 and 7 holds true for $\widehat{\mathbf{V}}$. Clearly, we have $\widehat{\mathcal{I}}_+ := \{(l, j) : \widehat{\mathrm{V}}_{lk} \neq 0 \text{ if } k = j \text{ or } (k, j) \in \widehat{\mathcal{E}}_+\} = \mathcal{I}_+$ whenever $\widehat{\mathcal{E}}_+ = \mathcal{E}_+$. Thus, we only need to show $\widehat{\mathcal{E}}_+ = \mathcal{E}_+$. We shall prove this by induction.

Given $\mathcal{G}^{\text{work}}$ and $\mathbf{V}^{\text{work}}$, the set of instruments on leaves is

$$\mathcal{B} = \Big\{ l : l \text{ minimizes } \|\widehat{\mathbf{V}}^{\text{work}}_{l,\cdot}\|_0 \text{ and } \|\widehat{\mathbf{V}}^{\text{work}}_{l,\cdot}\|_0 > 0 \Big\} = \Big\{ l : \|\mathbf{V}^{\text{work}}_{l,\cdot}\|_0 = 1 \Big\}.$$

Hence, X_l is an instrument of leaf variable Y_k in $\mathcal{G}^{\text{work}}$, when $l \in \mathcal{B}$ and k maximizes $|\mathrm{V}^{\text{work}}_{lk}|$. Hence, $\mathcal{L}$ is the index set of leaves in $\mathcal{G}^{\text{work}}$. By Proposition 7, all (k, j) such that Y_k is an unmediated parent of a peeled variable Y_j are in $\widehat{\mathcal{E}}_+$ and can be identified by Assumption 1B. In model (1), after removing $\boldsymbol{Y}_{\mathcal{L}}$, the local Markov property of the rest variables in $\mathcal{G}^{\text{work}}$ remain intact. Then we repeat the procedure until all primary variables are removed.

As a result, Algorithm 2 correctly identifies a subset of $\mathcal{E}_+$ that contains all edges from an unmediated parent, so it suffices to recover $\mathcal{E}_+$. Consequently, $\mathcal{E}_+$ can be reconstructed by Step 9 and $\mathcal{I}_+$ can be recovered by Step 10. This completes the proof of (B).

Lemma 22 *Let*

$$T_j = \frac{\mathbf{e}_j^\top (\mathbf{P}_{A_j} - \mathbf{P}_{B_j}) \mathbf{e}_j}{\mathbf{e}_j^\top (\mathbf{I} - \mathbf{P}_{A_j}) \mathbf{e}_j / (n - |A_j|)}, \quad T_j^* = \frac{(\mathbf{e}_j^*)^\top (\mathbf{P}^*_{A_j} - \mathbf{P}^*_{B_j}) \mathbf{e}_j^*}{(\mathbf{e}_j^*)^\top (\mathbf{I} - \mathbf{P}^*_{A_j}) \mathbf{e}_j^* / (n - |A_j|)}, \tag{31}$$

where $A_j = \text{PA}_{\mathcal{S}}(j) \cup \text{IN}_{\mathcal{S}}(j)$ *and* $B_j^\circ = (\text{PA}_{\mathcal{S}}(j) \cup \text{IN}_{\mathcal{S}}(j)) \setminus \text{D}_{\mathcal{S}}(j)$. *Then* $(T_1, \ldots, T_p)$ *are independent and* $(T_1^*, \ldots, T_p^*)$ *are conditionally independent given* $\mathbf{Z}$. *Moreover,*

$$T_j^* / (|A_j| - |B_j|) \mid \mathbf{Z} \sim T_j / (|A_j| - |B_j|) \sim F_{|A_j| - |B_j|, n - |A_j|}; \quad 1 \le j \le p,$$

where F_{d_1, d_2} *denotes the F-distribution with* d_1 *and* d_2 *degrees of freedom.*

Proof of Lemma 22

Let $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$ as in (13). Given data submatrix $\mathbf{Z}_{A_j}$,

$$\frac{\mathbf{e}_j^\top (\mathbf{P}_{A_j} - \mathbf{P}_{B_j}) \mathbf{e}_j}{\sigma_j^2} \mid \mathbf{Z}_{A_j} \sim \chi^2_{|A_j| - |B_j|}, \quad \frac{\mathbf{e}_j^\top (\mathbf{I} - \mathbf{P}_{A_j}) \mathbf{e}_j}{\sigma_j^2} \mid \mathbf{Z}_{A_j} \sim \chi^2_{n - |A_j|},$$

and they are independent, because $\mathbf{P}_{A_j} - \mathbf{P}_{B_j}$ and $\mathbf{I} - \mathbf{P}_{A_j}$ are orthonormal projection matrices, $\mathbf{e}_j \mid \mathbf{Z}_{A_j} \sim N(\mathbf{0}, \sigma_j^2 \mathbf{I}_n)$, and $(\mathbf{P}_{A_j} - \mathbf{P}_{B_j})(\mathbf{I} - \mathbf{P}_{A_j}) = \mathbf{0}$. Then, for any real number t, the characteristic function $t \mapsto \mathbb{E} \exp(\iota t T_j / (|A_j| - |B_j|))$ is (ι is the imaginary unit)

$$\mathbb{E}\left(\mathbb{E}\left(\exp(\iota t T_j / (|A_j| - |B_j|)) \mid \mathbf{Z}_{A_j} \right) \right) = \mathbb{E}\, \psi_{|A_j| - |B_j|, n - |A_j|}(t) = \psi_{|A_j| - |B_j|, n - |A_j|}(t),$$

where $\psi_{|A_j| - |B_j|, n - |A_j|}$ is the characteristic function of F-distribution with degrees of freedom $|A_j| - |B_j|, n - |A_j|$. Hence, $T_j / (|A_j| - |B_j|) \sim F_{|A_j| - |B_j|, n - |A_j|}$. Similarly, we also have $T_j^* / (|A_j| - |B_j|) \sim F_{|A_j| - |B_j|, n - |A_j|}$; $j = 1, \ldots, p$.

Next, we prove independence for $\boldsymbol{T}$ and $\boldsymbol{T}^*$ via a peeling argument. Let $\boldsymbol{t} = (t_1, \ldots, t_p)$. Let Y_j be a leaf node of the graph $\mathcal{G}$. Then $\boldsymbol{T}_{-j} \mid \mathbf{Y}_{-j}, \mathbf{X}$ is deterministic, where $\boldsymbol{T}_{-j}$ is the subvector of $\boldsymbol{T}$ with the jth component removed. The characteristic function of $\boldsymbol{t}^\top \boldsymbol{T}$ is

$$\begin{aligned} \mathbb{E} \exp(\iota \boldsymbol{t}^\top \boldsymbol{T}) &= \mathbb{E}\left(\mathbb{E}\left(\exp(\iota t_j T_j) \mid \mathbf{Y}_{-j}, \mathbf{X} \right) \exp(\iota \boldsymbol{t}_{-j}^\top \boldsymbol{T}_{-j}) \right) \\ &= \psi_{|A_j| - |B_j|, n - |A_j|}(t_j)\, \mathbb{E} \exp(\iota \boldsymbol{t}_{-j}^\top \boldsymbol{T}_{-j}), \end{aligned} \tag{32}$$

where $\psi_{|A_j|-|B_j|,n-|A_j|}$ is the characteristic function of the F-distribution with degrees of freedom $(|A_j|-|B_j|, n-|A_j|)$. Next, let $Y_{j'}$ be a leaf node of the graph $\mathcal{G}'$ and apply the law of iterated expectation again, where $\mathcal{G}'$ is the subgraph of $\mathcal{G}$ without node Y_j. Repeat this procedure and after p steps $\mathbb{E}\exp(\iota \boldsymbol{t}^\top \boldsymbol{T}) = \prod_{j=1}^p \psi_{|A_j|-|B_j|,n-|A_j|}(t_j)$, which implies $(T_1,\ldots,T_p)$ are independent. Similarly, $\boldsymbol{T}^*$ also has independent components given $\mathbf{Z}$ and has the same distribution as $\boldsymbol{T}$. This completes the proof.

B.5 Proof of Theorem 15

Let $\mathrm{Lr}(\mathcal{S}, \boldsymbol{\Sigma})$ denote the likelihood ratio given $\mathcal{S}$ and $\boldsymbol{\Sigma}$. Then $\mathrm{Lr} = \mathrm{Lr}(\widehat{\mathcal{S}}, \widehat{\boldsymbol{\Sigma}})$.

Proof of (A)

Without loss of generality, assume M is sufficiently large. For any real number x,

$$\begin{aligned} & |\,\mathbb{P}(\mathrm{Lr} \le x) - \mathbb{P}(\mathrm{Lr}(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}) \le x)| \\ \le\ & \mathbb{E}\,|\,\mathrm{I}(\mathrm{Lr} \le x) - \mathrm{I}(\mathrm{Lr}(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}) \le x)| \\ =\ & \mathbb{E}\,|\,\mathrm{I}(\mathrm{Lr} \le x, \widehat{\mathcal{S}} \ne \mathcal{S}) + \mathrm{I}(\mathrm{Lr}(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}) \le x)(\mathrm{I}(\widehat{\mathcal{S}} = \mathcal{S}) - 1)| \\ \le\ & 2\,\mathbb{P}(\widehat{\mathcal{S}} \ne \mathcal{S}). \end{aligned} \tag{33}$$

From (14), $\mathrm{Lr}^*(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}^*) = \sum_{\{j:\mathrm{D}_{\mathcal{S}}(j)\ne\emptyset\}} T_j^*$. By Lemma 22,

$$\mathbb{P}(\mathrm{Lr}(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}) \le x) = \mathbb{P}(\mathrm{Lr}^*(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}^*) \le x \mid \mathbf{Z}).$$

Note that $\mathrm{Lr}^* = \mathrm{Lr}^*(\widehat{\mathcal{S}}^*, \widehat{\boldsymbol{\Sigma}}^*)$. Then for any real number x,

$$\begin{aligned} & |\,\mathbb{P}(\mathrm{Lr}^* \le x \mid \mathbf{Z}) - \mathbb{P}(\mathrm{Lr}(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}) \le x)| \\ =\ & |\,\mathbb{P}(\mathrm{Lr}^* \le x, \widehat{\mathcal{S}}^* \ne \mathcal{S} \mid \mathbf{Z}) + \mathbb{P}(\mathrm{Lr}^* \le x, \widehat{\mathcal{S}}^* = \mathcal{S} \mid \mathbf{Z}) - \mathbb{P}(\mathrm{Lr}^*(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}^*) \le x \mid \mathbf{Z})| \\ \le\ & |\,\mathbb{P}(\mathrm{Lr}^* \le x, \widehat{\mathcal{S}}^* \ne \mathcal{S} \mid \mathbf{Z})| + |\,\mathbb{P}(\mathrm{Lr}^*(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}^*) \le x, \widehat{\mathcal{S}}^* \ne \mathcal{S} \mid \mathbf{Z})| \\ \le\ & 2\,\mathbb{P}(\widehat{\mathcal{S}}^* \ne \mathcal{S} \mid \mathbf{Z}). \end{aligned} \tag{34}$$

Since (33) and (34) hold uniformly in x, we have

$$\sup_{x\in\mathbb{R}} |\,\mathbb{P}(\mathrm{Lr} \le x) - \mathbb{P}(\mathrm{Lr}^* \le x \mid \mathbf{Z})| \le 2\,\mathbb{P}(\widehat{\mathcal{S}} \ne \mathcal{S}) + 2\,\mathbb{P}(\widehat{\mathcal{S}}^* \ne \mathcal{S} \mid \mathbf{Z}).$$

By Theorem 14, we have $\mathbb{P}(\widehat{\mathcal{S}} \ne \mathcal{S}) \to 0$, which holds uniformly for all $\boldsymbol{\theta}$ which satisfy the Assumptions 1-4. For $\mathbb{P}(\widehat{\mathcal{S}}^* \ne \mathcal{S} \mid \mathbf{Z})$, the error terms of (7) in the perturbed data $\mathbf{Z}^*$ are rescaled with $\Omega_{jj}^{-1} + \widehat{\sigma}_j^2 \le 2\Omega_{jj}^{-1}$. By Theorem 14, $\mathbb{P}(\widehat{\mathcal{S}}^* \ne \mathcal{S}) = \mathbb{E}\,\mathbb{P}(\widehat{\mathcal{S}}^* \ne \mathcal{S} \mid \mathbf{Z}) \to 0$, which implies $\mathbb{P}(\widehat{\mathcal{S}}^* \ne \mathcal{S} \mid \mathbf{Z}) \xrightarrow{p} 0$ as $n \to \infty$ by the Markov inequality. Consequently, $\sup_{x\in\mathbb{R}} |\,\mathbb{P}(\mathrm{Lr} \le x) - \mathbb{P}(\mathrm{Lr}^* \le x \mid \mathbf{Z})| \to 0$. For $|\mathcal{D}| = 0$, $\mathbb{P}(\mathrm{Lr} = 0) \to 1$, $\mathbb{P}(\mathrm{Lr}^* = 0 \mid \mathbf{Z}) \to 1$, and $\mathbb{P}(\mathrm{Pval} = 1) \to 1$. For $|\mathcal{D}| > 0$, $\mathbb{P}(\mathrm{Lr}^* \ge \mathrm{Lr} \mid \mathbf{Z}) \to \mathrm{Unif}(0,1)$ and $\mathbb{P}(\mathrm{Pval} < \alpha) \to \alpha$. This completes the proof of (A).

Proof of (B)

Let $\text{Pval}_k = M^{-1}\sum_{m=1}^{M} \text{I}(\text{Lr}^*_{k,m} \geq \text{Lr})$, the p-value of sub-hypothesis $\text{H}_{0,k}$. For $|\mathcal{D}| < |\mathcal{H}|$, there exists an edge $(i_k, j_k) \in \mathcal{H}$ but $(i_k, j_k) \notin \mathcal{D}$. Then by (A), $\mathbb{P}(\text{Pval} = \text{Pval}_k = 1) \to 1$. For $|\mathcal{D}| = |\mathcal{H}|$, note that as $n, M \to \infty$,

$$\mathbb{P}\left(\text{Pval} < \alpha\right) = \mathbb{P}\left(\text{Pval}_1 < \alpha, \ldots, \text{Pval}_{|\mathcal{H}|} < \alpha\right) \leq \mathbb{P}\left(\text{Pval}_1 < \alpha\right) \to \alpha.$$

Now, define a sequence $\{\mathbf{U}^{(r)}_{\mathcal{H}}\}_{r\geq 1}$ such that $\mathbf{U}^{(r)}_{(i_1,j_1)} = 0$ and $\min_{2\leq k\leq |\mathcal{H}|} |\mathbf{U}^{(r)}_{(i_k,j_k)}| \geq c > 0$. Thus, $\{\mathbf{U}^{(r)}_{\mathcal{H}}\}_{r\geq 1}$ satisfy H_0. By Proposition 18, $\text{Pval}_k \xrightarrow{p} 0$ for $k \geq 2$ as $r \to \infty$. Hence,

$$\limsup_{\substack{n\to\infty \\ \mathbf{U}^{(r)}_{\mathcal{H}} \text{ satisfies } \text{H}_0}} \mathbb{P}_{\boldsymbol{\theta}^{(r)}}\left(\text{Pval} < \alpha\right) \geq \sup_r \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(r)}}\left(\text{Pval} < \alpha\right) = \alpha.$$

This completes the proof.

B.6 Proof of Proposition 16

Since $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$, it suffices to consider $\text{Lr}(\mathcal{S}, \widehat{\boldsymbol{\Sigma}})$. For $|\mathcal{D}| = 0$, we have $\mathbb{P}(\text{Lr} = 0) \to 1$. Now, assume $|\mathcal{D}| > 0$. Then

$$2\text{Lr}(\mathcal{S}, \widehat{\boldsymbol{\Sigma}}) = \underbrace{\sum_{\{j:\text{D}_{\mathcal{S}}(j)\neq\emptyset\}} \frac{\mathbf{e}_j^\top(\mathbf{P}_{A_j} - \mathbf{P}_{B_j})\mathbf{e}_j}{\sigma_j^2}}_{R_1} + \underbrace{\sum_{\{j:\text{D}_{\mathcal{S}}(j)\neq\emptyset\}} \left(\frac{\sigma_j^2}{\widehat{\sigma}_j^2} - 1\right) \frac{\mathbf{e}_j^\top(\mathbf{P}_{A_j} - \mathbf{P}_{B_j})\mathbf{e}_j}{\sigma_j^2}}_{R_2}.$$

To derive the asymptotic distribution of R_1, we apply the strategy with law of iterated expectation as in the proof of Lemma 22. Then we have $\left\{\mathbf{e}_j^\top(\mathbf{P}_{A_j} - \mathbf{P}_{B_j})\mathbf{e}_j/\sigma_j^2\right\}_{\{j:\text{D}_{\mathcal{S}}(j)\neq\emptyset\}}$ are independent. Therefore, $R_1 \sim \chi^2_{|\mathcal{D}|}$. To bound R_2, we apply Lemma 1 of Laurent and Massart (2000). By Assumption 5,

$$\mathbb{P}\left(\max_{\{j:\text{D}_{\mathcal{S}}(j)\neq\emptyset\}} \left|\frac{\widehat{\sigma}_j^2}{\sigma_j^2} - 1\right| \geq 4\sqrt{\frac{\log|\mathcal{D}|}{(1-\rho)n}} + 8\frac{\log|\mathcal{D}|}{(1-\rho)n}\right) \leq 2\exp(-\log|\mathcal{D}|).$$

Hence, $\max_{\{j:\text{D}_{\mathcal{S}}(j)\neq\emptyset\}} \left|\frac{\sigma_j^2}{\widehat{\sigma}_j^2} - 1\right| \leq 8\sqrt{\log(|\mathcal{D}|)/(1-\rho)n}$ with probability tending one. Thus

$$|R_2| \leq |R_1| \max_{\{j:\text{D}_{\mathcal{S}}(j)\neq\emptyset\}} \left|\frac{\sigma_j^2}{\widehat{\sigma}_j^2} - 1\right| \leq O_{\mathbb{P}}\left(|\mathcal{D}|\sqrt{\frac{\log|\mathcal{D}|}{n}}\right).$$

Consequently, the desired result follows.

B.7 Proof of Proposition 18

Let $\boldsymbol{\theta}^{(n)} = (\mathbf{U}^\circ + \boldsymbol{\Delta}, \mathbf{W}^\circ)$. Then

$$L(\boldsymbol{\theta}^{(n)}, \boldsymbol{\Sigma}) - L(\boldsymbol{\theta}^\circ, \boldsymbol{\Sigma}) = \sum_{\{j:\text{D}_{\mathcal{S}}(j)\neq\emptyset\}} \left(\sqrt{n}\boldsymbol{\eta}_j^\top \boldsymbol{\Delta}_{\cdot j} - \frac{1}{2}\sqrt{n}\boldsymbol{\Delta}_{\cdot j}\left(n^{-1}\mathbf{Y}^\top_{\text{D}_{\mathcal{S}}(j)}\mathbf{Y}_{\text{D}_{\mathcal{S}}(j)}\right)\sqrt{n}\boldsymbol{\Delta}_{\cdot j}\right),$$

where $\boldsymbol{\eta}_j = (n^{-1/2}\mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)}^\top \mathbf{e}_j, \mathbf{0}_{|\mathrm{D}_{\mathcal{S}}(j)^c|})$ is a p-vector. It suffices to consider $\mathrm{Lr}(\mathcal{S}, \widehat{\boldsymbol{\Sigma}})$, since $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$. Under null hypothesis H_0, $2\mathrm{Lr} = \sum_{\{j:\mathrm{D}_{\mathcal{S}}(j)\neq\emptyset\}} T_j$, where by Proposition 16 we have $T_j = \sum_{r=1}^{|\mathrm{D}_{\mathcal{S}}(j)|}(\mathbf{q}_{j,r}^\top \mathbf{e}_j)^2 + o_{\mathbb{P}}(1)$ with $\mathbf{P}_{A_j} - \mathbf{P}_{B_j} = \sum_{r=1}^{|\mathrm{D}_{\mathcal{S}}(j)|} \mathbf{q}_{j,r}\mathbf{q}_{j,r}^\top$. Letting $\mathbf{Q}_j = \big(\mathbf{q}_{j,1}, \ldots, \mathbf{q}_{j,|\mathrm{D}_{\mathcal{S}}(j)|}\big)$ be an $n \times |\mathrm{D}_{\mathcal{S}}(j)|$ matrix, then

$$\begin{pmatrix} \mathbf{Q}_j^\top \mathbf{e}_j \\ \boldsymbol{\eta}_j \end{pmatrix} \Bigg| \, \mathbf{Y}_{\mathrm{PA}_{\mathcal{S}}(j)}, \mathbf{X} \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{I}_{|\mathrm{D}_{\mathcal{S}}(j)|} & n^{-1/2}\mathbf{Q}_j^\top \mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)} \\ n^{-1/2}\mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)}^\top \mathbf{Q}_j & n^{-1}\mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)}^\top \mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)} \end{pmatrix}\right),$$

$$\mathbf{Q}_j^\top \mathbf{e}_j \mid \boldsymbol{\eta}_j, \mathbf{Y}_{\mathrm{PA}_{\mathcal{S}}(j)}, \mathbf{X} \sim N\Big(n^{-1/2}\mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)}^\top \mathbf{Q}_j \boldsymbol{\eta}_j, \mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)}^\top (\mathbf{I}_n - \mathbf{Q}_j\mathbf{Q}_j^\top)\mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(j)}\Big).$$

Next, let Y_k be a leaf node of the graph $\mathcal{G}$. For fixed $|\mathcal{D}| > 0$, after change of measure,

$$\begin{aligned}
&\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta}) \\
&\geq \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}^\circ}\left(\mathrm{I}\left(\sum_{j=1}^p \|\mathbf{Q}^\top \mathbf{e}_j\|_2^2 > \chi^2_{|\mathcal{D}|,1-\alpha}\right)\exp\Big(L(\boldsymbol{\theta}^{(n)}, \boldsymbol{\Sigma}) - L(\boldsymbol{\theta}^\circ, \boldsymbol{\Sigma})\Big)\right) \\
&= \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}^\circ}\, \mathbb{E}\left(\mathrm{I}\left(\sum_{j=1}^p \|\mathbf{Q}^\top \mathbf{e}_j\|_2^2 > \chi^2_{|\mathcal{D}|,1-\alpha}\right)\exp\Big(L(\boldsymbol{\theta}^{(n)}, \boldsymbol{\Sigma}) - L(\boldsymbol{\theta}^\circ, \boldsymbol{\Sigma})\Big) \,\middle|\, \mathbf{Y}_{-k}, \mathbf{X}\right) \\
&= \liminf_{n\to\infty} \mathbb{E}\left(\mathbb{P}\left(\|\boldsymbol{Z}_k + \mathbf{Q}_k^\top \mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(k)}\boldsymbol{\Delta}_{\cdot k}\|_2^2 + \sum_{j\neq k}\|\mathbf{Q}_j^\top \mathbf{e}_j\|_2^2 > \chi^2_{|\mathcal{D}|,1-\alpha} \,\middle|\, \mathbf{Y}_{-k}, \mathbf{X}\right)\right),
\end{aligned}$$

where $\boldsymbol{Z}_k \sim N(\mathbf{0}, \mathbf{I}_{|\mathrm{D}_{\mathcal{S}}(k)|})$. By Lemma 3 of Li et al. (2020), $\|\mathbf{Q}_k^\top \mathbf{Y}_{\mathrm{D}_{\mathcal{S}}(k)}\boldsymbol{\Delta}_{\cdot k}\|_2^2 \geq c_3\|\boldsymbol{\Delta}_{\cdot k}\|_2^2$ with probability tending to one. Therefore, a peeling argument yields

$$\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta}) \geq \mathbb{P}\Big(\|\boldsymbol{Z} + c_3\boldsymbol{\Delta}\|_2^2 \geq \chi^2_{|\mathcal{D}|,1-\alpha}\Big),$$

where $\boldsymbol{Z} \sim N(\mathbf{0}, \mathbf{I}_{|\mathcal{D}|})$. Similarly, as $|\mathcal{D}| \to \infty$,

$$\beta(\boldsymbol{\theta}^\circ, \boldsymbol{H}) \geq \mathbb{P}\Big(\|\boldsymbol{Z} + c_3\boldsymbol{\Delta}\|_2^2 > \sqrt{2|\mathcal{D}|}z_{1-\alpha} + |\mathcal{D}|\Big) \to \mathbb{P}\Big(Z > z_{1-\alpha} - c_3\|\boldsymbol{\Delta}\|_2^2/\sqrt{2|\mathcal{D}|}\Big).$$

This completes the proof.

B.8 Proof of Proposition 19

By Proposition 18, for fixed $|\mathcal{D}| = |\mathcal{H}| > 0$, $\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta})$ equals to

$$\mathbb{P}(\mathrm{Pval}_1 < \alpha, \ldots, \mathrm{Pval}_{|\mathcal{H}|} < \alpha) \geq 1 - \sum_{k=1}^{|\mathcal{H}|}\mathbb{P}(\mathrm{Pval}_k \leq \alpha) \geq 1 - |\mathcal{H}|\,\mathbb{P}\left(Z_1^2 \leq \chi^2_{1,1-\alpha}\right),$$

where $Z_1 \sim N(\delta/\max_{1\leq j\leq p}\Omega_{jj}, 1)$. Then $\mathbb{P}(\widetilde{Z}^2 \leq x) \leq \mathbb{P}(\widetilde{Z} \leq -\mu + \sqrt{x}) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-(\mu-\sqrt{x})^2/2}}{\mu-\sqrt{x}}$ for $\widetilde{Z} \sim N(\mu, 1)$ and any $0 \leq x < \mu^2$ and $\mu > 0$ (Small, 2010). Hence, as $|\mathcal{D}| = |\mathcal{H}| \to \infty$,

$$\begin{aligned}
\beta(\boldsymbol{\theta}^\circ, \boldsymbol{\Delta}) &\geq 1 - |\mathcal{H}|\,\mathbb{P}\left(Z_1^2 \leq \chi^2_{1,1-\alpha}\right) \\
&\geq 1 - \frac{|\mathcal{H}|}{\sqrt{2\pi}}e^{-\left(\delta\sqrt{\log|\mathcal{H}|}/\max_{1\leq j\leq p}\Omega_{jj} - \sqrt{\chi^2_{1,1-\alpha}}\right)^2/2} \to 1.
\end{aligned}$$

This completes the proof.

B.9 Proof of Theorem 20

The proof proceeds in two steps to show that
(A) $\{l : |\mathrm{V}_{lj}| \geq \tau_j^*\} \subseteq \{l : \widehat{\mathrm{V}}_{lj} \neq 0\} \subseteq \{l : \mathrm{V}_{lj} \neq 0\}$ for $1 \leq j \leq p$ almost surely, and
(B) $\widehat{\mathcal{G}}_+ = \mathcal{G}_+$ when $\widehat{\mathbf{V}}$ satisfies the property in (A).

Proof of (A)

This proof is similar to that of Theorem 14 part (A). To proceed, let $A_j^\circ = \Big\{l : \mathrm{V}_{lj} \neq 0\Big\}$, $A_j^* = \Big\{l : |\mathrm{V}_{lj}| > \tau_j^*\Big\}$, and $A_j^{[t]} = \Big\{l : |\widetilde{\mathrm{V}}_{lj}^{[t]}| > \tau_j\Big\}$. Then we define the false negative set $\mathrm{FN}_j^{[t]} = A_j^* \setminus A_j^{[t]}$ and the false positive set $\mathrm{FP}_j^{[t]} = A_j^{[t]} \setminus A_j^\circ$ for $t \geq 0$. Consider

$$\mathscr{E}_j = \Big\{\|\mathbf{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty \leq 0.5\gamma_j\tau_j\Big\} \cap \Big\{\|\widehat{\mathbf{V}}_{\cdot j}^\circ - \mathbf{V}_{\cdot j}\|_\infty \leq 0.5\tau_j\Big\},$$

where $\widehat{\boldsymbol{\xi}}_j = \mathbf{Y}_j - \mathbf{X}\widehat{\mathbf{V}}_{\cdot j}^\circ$ for $1 \leq j \leq p$. Again, we shall show that $\mathrm{FN}_j^{[t]}$ and $\mathrm{FP}_j^{[t]}$ are eventually empty on event $\mathscr{E}_j$ which has a probability tending to one.

Assume $|A_j^\circ \cup A_j^{[t-1]}| \leq 2\kappa_{\max}^\circ$. Then (28), (29), and (30) remain true on $\mathscr{E}_j$. As a result, $|A_j^\circ \cup A_j^{[t]}| \leq 2\kappa_{\max}^\circ$ on $\mathscr{E}_j$ for $t \geq 1$.

To estimate the number of iterations required for termination, it suffices to prove that $|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}| < 1$ eventually. Suppose $|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}| \geq 1$. For any $l \in \mathrm{FN}_j^{[t]} \cup \mathrm{FP}_j^{[t]}$, by Assumption 5,

$$|\widetilde{\mathrm{V}}_{lj}^{[t]} - \widehat{\mathrm{V}}_{lj}^\circ| \geq |\widetilde{\mathrm{V}}_{lj}^{[t]} - V_{lj}| - |\widehat{\mathrm{V}}_{lj}^\circ - \mathrm{V}_{lj}| \geq \min(\tau_j^* - \tau_j, \tau_j),$$

so $\sqrt{|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}|} \leq \|\widetilde{\mathbf{V}}_{\cdot j}^{[t]} - \widehat{\mathbf{V}}_{\cdot j}^\circ\|_2 / \min(\tau_j^* - \tau_j, \tau_j)$. Letting $\kappa_j^* = |A_j^*|$, by (30) and the Cauchy-Schwarz inequality,

$$c_1\|\widetilde{\mathbf{V}}_{\cdot j}^{[t]} - \widehat{\mathbf{V}}_{\cdot j}^\circ\|_2 \leq 1.5\gamma_j\tau_j\sqrt{|A_j^\circ \triangle A_j^{[t-1]}|} \leq 1.5\gamma_j\tau_j\sqrt{|\mathrm{FN}_j^{[t-1]}| + |\mathrm{FP}_j^{[t-1]}| + (\kappa_j^\circ - \kappa_j^*)}.$$

By the conditions for (τ_j, γ_j) in Theorem 20, we have

$$\begin{aligned}
\sqrt{|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}|} &\leq \frac{\|\widetilde{\mathbf{V}}_{\cdot j}^{[t]} - \widehat{\mathbf{V}}_{\cdot j}^\circ\|_2}{\min(\tau_j^* - \tau_j, \tau_j)} \\
&\leq \frac{1.5\gamma_j\tau_j}{c_1 \min(\tau_j^* - \tau_j, \tau_j)}\Big(\sqrt{|\mathrm{FN}_j^{[t-1]}| + |\mathrm{FP}_j^{[t-1]}|} + \sqrt{\kappa_j^\circ - \kappa_j^*}\Big) \\
&\leq 0.5\sqrt{|\mathrm{FN}_j^{[t-1]}| + |\mathrm{FP}_j^{[t-1]}|} + 0.25.
\end{aligned}$$

Hence, $\sqrt{|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}|} \leq (1/2)^t\sqrt{|A_j^\circ| + |A_j^{[0]}|} + 0.5$. For $t \geq 1 + \lceil \log \kappa_j^\circ / \log 4 \rceil$, we have $|\mathrm{FN}_j^{[t]}| + |\mathrm{FP}_j^{[t]}| < 1$ implying that $\mathrm{FN}_j^{[t]} = \emptyset$ and $\mathrm{FP}_j^{[t]} = \emptyset$.

Under the conditions for (τ_j, γ_j) in Theorem 20, using the same argument in Proof of Theorem 14 we have

$$\mathbb{P}\left(\Big\{l : |\mathrm{V}_{lj}| > \tau_j^*\Big\} \subseteq \Big\{l : |\widetilde{\mathrm{V}}_{lj}| > \tau_j\Big\} \subseteq \Big\{l : \mathrm{V}_{lj} \neq 0\Big\};\ 1 \leq j \leq p\right) \geq 1 - \sqrt{\frac{2}{\pi}}pq^{-3}n^{-4}.$$

By Borel-Cantelli lemma, almost surely we have

$$\Big\{l : |\mathrm{V}_{lj}| > \tau_j^*\Big\} \subseteq \Big\{l : |\widetilde{\mathrm{V}}_{lj}| > \tau_j\Big\} \subseteq \Big\{l : \mathrm{V}_{lj} \neq 0\Big\}, \quad j = 1, \ldots, p,$$

when n is sufficiently large. Thus, we have $\{l : \widehat{\mathrm{V}}_{lj} \neq 0\} = \{l : |\widetilde{\mathrm{V}}_{lj}| > \tau_j\}$ whenever $\kappa_j = |\{l : \widetilde{\mathrm{V}}_{lj} > \tau_j\}|$; $j = 1, \ldots, p$. The desired result follows.

Proof of (B)

This is the same as the Proof of Theorem 14 part (B).

B.10 Proof of Theorem 21

By Theorem 14, it suffices to consider the event $\{\widehat{\mathcal{G}}_+ = \mathcal{G}_+\}$. Then with $\mathbf{X}$ being replaced by $(\mathbf{Y}_{\cdot,\mathrm{AN}_{\mathcal{G}_+}(j)}, \mathbf{X}_{\cdot,\mathrm{IN}_{\mathcal{G}_+}(j)})$, this proof is almost the same as that of Theorem 14 part (A) and thus is omitted.

Appendix C. Supplements for Simulations

C.1 Implementation Details

DP-LR and LR:

- For Algorithm 3, we fix the Monte Carlo sample size $M = 500$. Our limited numerical experience suggests that this choice appears adequate for our purpose.
- For Algorithms 1 and 2, we choose $\tau_j \in \{0.05, 0.1, 0.15\}$ and $\gamma_j = \exp(\gamma_j')$ with

  $$\gamma_j' \in \Big\{ \log(\max_{l,j} |\mathbf{X}_{\cdot l}^\top \mathbf{Y}_{\cdot j}|), \ldots, 0.05 \log(\max_{l,j} |\mathbf{X}_{\cdot l}^\top \mathbf{Y}_{\cdot j}|) \Big\} \quad \text{(100 equally spaced values)}.$$

  Then BIC is used to estimate tuning parameters $\kappa_j \in \{1, \ldots, 30\}$; $j = 1, \ldots, p$.

2SPLS (Chen et al., 2018):

- We use the R package `BigSEM` for 2SPLS. In our experiments, the λ sequence of 2SPLS (Chen et al., 2018) is set in the same way as the γ sequence of the proposed methods. We use the default settings for other parameters.

C.2 Additional Simulations

This section supplements Section 5.2 by including additional numerical experiments.

C.2.1 Random graphs with different sparsity

We examine the proposed method for structure learning of DAGs with different sparsity. For $\mathbf{U}$, the upper off-diagonal entries U_{kj}; $k < j$ are sampled independently from $\{0, 1\}$ according to Bernoulli(s/p); $s = 1, 2, 3, 4$, while other entries are zero. The rest of the settings remain the same as in Section 5.2. Figure 11 displays the results. As expected, the performance improves when the sample size n increases and the DAG becomes sparser (controlled by s).

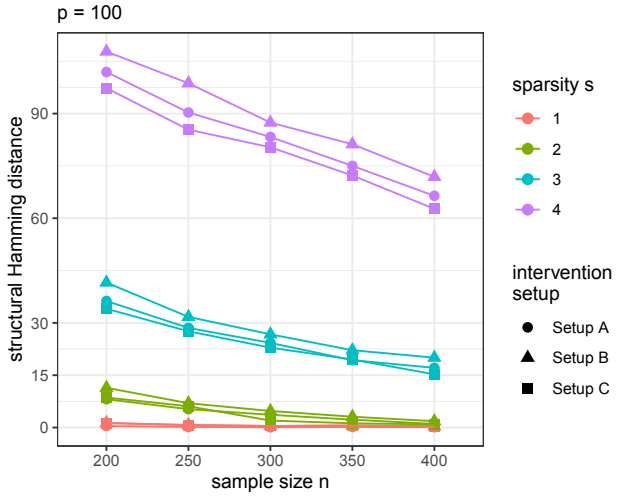


Figure 11: SHDs for the reconstructed DAG by the peeling algorithm.

C.2.2 SHD transition curves of structure learning

We consider different sample sizes $n = 50, 100, 150, 200$ to further examine how the proposed method depends on n. Figure 12 displays the SHD transition curves of the peeling algorithm.

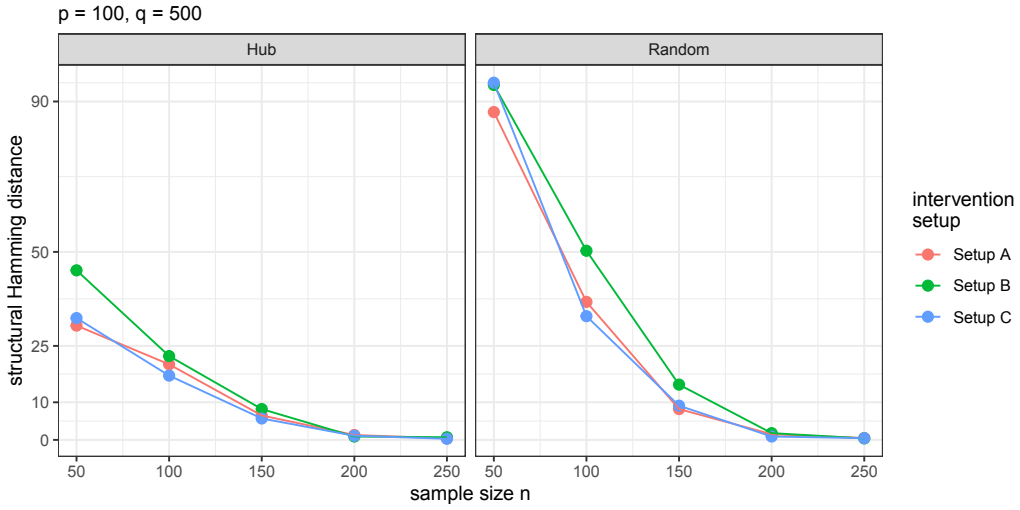


Figure 12: SHDs for the reconstructed DAG by the peeling algorithm. The experiment settings are the same as the ones in Section 5.2.

C.2.3 Structure learning with different numbers of interventions

Finally, we investigate how the number of interventions q affects the learning outcomes. Figure 13 displays the results when $q = 500, 1000, 1500$. It suggests that the proposed method

performs reasonably well at a moderate sample size when many unknown interventions are present.

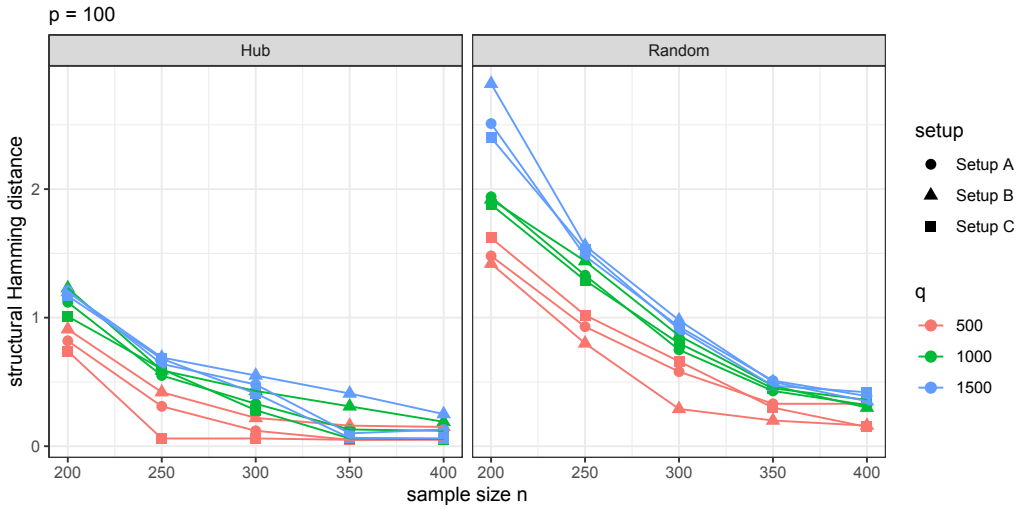


Figure 13: SHDs for the reconstructed DAG by the peeling algorithm. The experiment settings are the same as the ones in Section 5.2.

References

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Til Ole Bergmann and Gesa Hartwigsen. Inferring causality from noninvasive brain stimulation in cognitive neuroscience. *Journal of Cognitive Neuroscience*, 33(2):195–225, 2021.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253, 2011.

Leo Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419): 738–754, 1992.

Brielin C Brown and David A Knowles. Phenome-scale causal network discovery with bidirectional mediated Mendelian randomization. *bioRxiv*, 2020.

Guojun Bu. Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. *Nature Reviews Neuroscience*, 10(5):333–344, 2009.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Chen Chen, Min Ren, Min Zhang, and Dabao Zhang. A two-stage penalized least squares method for constructing large systems of structural equations. *Journal of Machine Learning Research*, 19(1):40–73, 2018.

Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 107–114. PMLR, 2007.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849, 2014.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Nir Friedman and Daphne Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1):95–125, 2003.

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

Moritz Grosse-Wentrup, Dominik Janzing, Markus Siegel, and Bernhard Schölkopf. Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage*, 125:825–833, 2016.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

Aimee L Jackson, Steven R Bartz, Janell Schelter, Sumire V Kobayashi, Julja Burchard, Mao Mao, Bin Li, Guy Cavet, and Peter S Linsley. Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, 21(6):635–637, 2003.

Jana Janková and Sara van de Geer. Inference in high-dimensional graphical models. In *Handbook of Graphical Models*, pages 325–350. CRC Press, 2018.

TCW Julia and Alison M Goate. Genetics of β-amyloid precursor protein in Alzheimer's disease. *Cold Spring Harbor Perspectives in Medicine*, 7(6), 2017.

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

Raymond J Kelleher III and Jie Shen. Presenilin-1 mutations and Alzheimer's disease. *Proceedings of the National Academy of Sciences*, 114(4):629–631, 2017.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. The big data bootstrap. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1787–1794, 2012.

Anna Kremer, Justin V Louis, Tomasz Jaworski, and Fred Van Leuven. GSK3 and Alzheimer's disease: facts and fiction. *Frontiers in Molecular Neuroscience*, 4:17, 2011.

Meghana M Kulkarni, Matthew Booker, Serena J Silver, Adam Friedman, Pengyu Hong, Norbert Perrimon, and Bernard Mathey-Prevot. Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays. *Nature Methods*, 3(10): 833–838, 2006.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Chunlin Li, Xiaotong Shen, and Wei Pan. Likelihood ratio tests for a large directed acyclic graph. *Journal of the American Statistical Association*, 115(531):1304–1319, 2020.

Chunlin Li, Xiaotong Shen, and Wei Pan. Nonlinear causal discovery with confounders. *Journal of the American Statistical Association*, pages 1–32, 2023.

Zhian Liu, Ming Zhang, Gongcheng Xu, Congcong Huo, Qitao Tan, Zengyong Li, and Quan Yuan. Effective connectivity analysis of the brain network in drivers during actual driving using near-infrared spectroscopy. *Frontiers in Behavioral Neuroscience*, 11:211, 2017.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

Po-Ling Loh and Martin J Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Ruiyan Luo and Hongyu Zhao. Bayesian hierarchical modeling for signaling pathway inference from single cell interventional data. *The Annals of Applied Statistics*, 5(2A):725–745, 2011.

Toshifumi Matsui, Martin Ingelsson, Hiroaki Fukumoto, Karunya Ramasamy, Hisatomo Kowa, Matthew P Frosch, Michael C Irizarry, and Bradley T Hyman. Expression of APP pathway mRNAs and proteins in Alzheimer's disease. *Brain Research*, 1161:116–123, 2007.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Aaron J Molstad, Wei Sun, and Li Hsu. A covariance-enhanced approach to multitissue joint eQTL mapping with application to transcriptome-wide association studies. *The Annals of Applied Statistics*, 15(2):998–1016, 2021.

Michael Murray. Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives*, 20(4):111–132, 2006.

Chris J Oates, Jim Q Smith, and Sach Mukherjee. Estimating causal structure using conditional DAG models. *Journal of Machine Learning Research*, 17(1):1880–1903, 2016.

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

Judea Pearl. *Causality*. Cambridge University Press, 2009.

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

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1): 2009–2053, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Goutham Rajendran, Bohdan Kivva, Ming Gao, and Bryon Aragam. Structure learning in polynomial time: Greedy algorithms, Bregman information, and exponential families. In *Advances in Neural Information Processing Systems*, volume 34, pages 18660–18672, 2021.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

Dominik Rothenhäusler, Peter Bühlmann, and Nicolai Meinshausen. Causal Dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3):1688–1722, 2019.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Xiaotong Shen and Jianming Ye. Adaptive model selection. *Journal of the American Statistical Association*, 97(457):210–221, 2002.

Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

Xiaotong Shen, Wei Pan, Yunzhang Zhu, and Hui Zhou. On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5): 807–832, 2013.

Chengchun Shi, Rui Song, Zhao Chen, and Runze Li. Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics*, 47(5):2671–2703, 2019.

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

Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

Christopher G Small. *Expansions and Asymptotics for Statistics*. Chapman and Hall/CRC, 2010.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

Joseph H Su, Ming Zhao, Aileen J Anderson, Anu Srinivasan, and Carl W Cotman. Activated caspase-3 expression in Alzheimer's and aged control brain: correlation with Alzheimer pathology. *Brain Research*, 898(2):350–357, 2001.

Alexander Teumer. Common methods for performing Mendelian randomization. *Frontiers in Cardiovascular Medicine*, 5:51, 2018.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

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

Jussi Viinikka, Antti Hyttinen, Johan Pensar, and Mikko Koivisto. Towards scalable Bayesian learning of causal DAGs. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6584–6594, 2020.

Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

Yiping Yuan, Xiaotong Shen, Wei Pan, and Zizhuo Wang. Constrained likelihood for reconstructing a directed acyclic Gaussian graph. *Biometrika*, 106(1):109–125, 2019.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948. PMLR, 2014.

Tuo Zhao, Han Liu, and Tong Zhang. Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics*, 46(1):180–218, 2018.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9492–9503, 2018.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association*, 115(529):217–230, 2020.