Inference for a Large Directed Acyclic Graph with Unspecified Interventions

Chunlin Li*,
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

Xiaotong Shen,
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

Wei Pan
Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

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 to identify 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 through the regional effects of neurons (Liu et al., 2017). In such a situation, a Gaussian directed acyclic graph (DAG) 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); see Heinze-Deml et al. (2018) for a review. For inference, 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 the strength of a single directed relation, and Li et al. (2020) derives a constrained likelihood ratio test for multiple directed relations.

Despite progress, challenges remain. First, inferring directed relations requires to identify 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 $O(p \times \log \kappa_{\text{max}} \times (q^3 + nq^2))$ operations almost surely, where $p$, $q$ are the numbers of primary and intervention variables, $n$ is the sample size, and $\kappa_{\text{max}}$ 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) $Y = (Y_1, \ldots, Y_p)^T$, consider a structural equation model with $q$ additive interventions:
\[ Y = U^T Y + W^T X + \epsilon, \quad \epsilon \sim N(0, \Sigma), \quad \Sigma = \text{Diag}(\sigma_1^2, \ldots, \sigma_p^2), \]  

where \( X = (X_1, \ldots, X_q)^T \) is a vector of additive intervention variables, \( U \in \mathbb{R}^{p \times p} \) and \( W \in \mathbb{R}^{q \times p} \) are unknown coefficient matrices, and \( \epsilon = (\epsilon_1, \ldots, \epsilon_p)^T \) is a vector of random errors with \( \sigma_j^2 > 0; j = 1, \ldots, p \). In (1), \( \epsilon \) is independent of \( X \) but the components of \( X \) can be dependent. The matrix \( U \) specifies the directed relations among \( Y \), where \( U_{ij} \neq 0 \) if \( Y_i \) is a direct cause of \( Y_j \) denoted by \( Y_i \rightarrow Y_j \) and \( Y_i \) is called a parent of \( Y_j \) or \( Y_j \) a child of \( Y_i \). Thus, \( U \) 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 \( W \) specifies the targets and strengths of interventions, where \( W_{lj} \neq 0 \) indicates \( X_l \) intervenes on \( Y_j \), denoted by \( X_l \rightarrow Y_j \). 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 \( \mathcal{G} = (Y, X; \mathcal{E}, \mathcal{I}) \) with primary variables \( Y \), intervention variables \( X \), primary edges \( \mathcal{E} = \{k,j): U_{ij} \neq 0\} \), and intervention edges \( \mathcal{I} = \{l,j): W_{lj} \neq 0\} \). To facilitate discussion, we introduce some concepts and notations for \( \mathcal{G} \). If there is a directed path \( Y_k \rightarrow \cdots \rightarrow Y_j \) in \( \mathcal{G} \), \( Y_k \) is an ancestor of \( Y_j \) or \( Y_j \) is a descendant of \( Y_k \). If \( Y_k \rightarrow Y_j \) 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 \( \text{PA}_\mathcal{G}(j) = \{k: Y_k \rightarrow Y_j\}, \text{AN}_\mathcal{G}(j) = \{k: Y_k \rightarrow \cdots \rightarrow Y_j\}, \) and \( \text{IN}_\mathcal{G}(j) = \{l: X_l \rightarrow Y_j\} \) be the parent, ancestor, and intervention sets of \( Y_j \) in DAG \( \mathcal{G} \), respectively.

### 2.1 Identifiability and instruments

Model (1) is generally non-identifiable without interventions (\( W = 0 \)) 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 \( \epsilon \) 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 $Y_1 \rightarrow Y_2$ and $Y_2 \rightarrow Y_1$.

**Remark 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.

1. **(1A)** $XX^\top$ is positive definite.
2. **(1B)** $\text{Cov}(Y_j, X_l | X_{(1, \ldots, q)} \setminus l) \neq 0$ if $X_l$ intervenes on any unmediated parent of $Y_j$.
3. **(1C)** Each primary variable is intervened by at least one instrument.

Assumption 1A imposes mild distributional restrictions on $X$, 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 1** Under Assumption 1, $(U, W, \Sigma)$ in model (1) are identifiable from the distribution of $(Y, X)$.

Proposition 1 (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, it is worth noting that 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 $Y_k \rightarrow Y_j$ 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,
\[ H_0: U_{ij} = 0; \text{for all} \ (k, j) \in \mathcal{H} \quad \text{versus} \quad H_a: U_{ij} \neq 0 \text{for some} \ (k, j) \in \mathcal{H}; \quad (2) \]

for simultaneous testing of directed pathways,

\[ H_0: U_{ij} = 0; \text{for some} \ (k, j) \in \mathcal{H} \quad \text{versus} \quad H_a: U_{ij} \neq 0 \text{for all} \ (k, j) \in \mathcal{H}, \quad (3) \]

where \((U_{\mathcal{H}}, W, \Sigma)\) 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

\[ H_{0, \nu}: U_{k_\nu j_\nu} = 0, \quad \text{versus} \quad H_{a, \nu}: U_{k_\nu j_\nu} \neq 0; \quad \nu = 1, \ldots, |\mathcal{H}|, \]

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 \((U, W)\) as well as identify the nonzero elements in \(U\) to recover the directed edges among the primary variables \(Y\) in \(\mathcal{H}\).

### 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 \(Y_{n \times p} = Y_1, \ldots, Y_n\) and \(X_{n \times q} = X_1, \ldots, X_n\) are given, where the rows \(\{(Y_i^\top, X_i^\top)\}_{1 \leq i \leq n}\) are independently sampled from (1). Then the log-likelihood is (up to a constant)

\[ L(\theta, \Sigma) = -\frac{1}{2} \sum_{i=1}^n \left\| \Sigma^{-1/2}(I - U^\top)Y_i - W^\top X_i \right\|_2^2 - n\log \det(\Sigma), \quad (4) \]

where \(\theta = (U, W), \Sigma = \text{Diag}(\sigma_1^2, \ldots, \sigma_p^2)\), and \(U\) 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 2 (Ancestral relation graph)**

**A.** A graph \(\mathcal{M} = (Y, X; \mathcal{A}, \emptyset)\) is an ARG if it is acyclic and \(\mathcal{A} = \{(k, j): k \in AN_{\mathcal{A}}(j)\}\).
B. Given DAG $\mathcal{G} = (\mathbf{Y}, \mathbf{X}; \mathcal{E}, \mathcal{I})$, its ARG is defined as $\mathcal{G}^+ = (\mathbf{Y}, \mathbf{X}; \mathcal{E}^+, \mathcal{I}^+)$, where

$$\mathcal{E}^+ = \{(k, j): k \in AN_{\mathcal{G}}(j)\}, \quad \mathcal{I}^+ = \bigcup_{k \in AN_{\mathcal{G}}(j) \cup \{j\}} IN_{\mathcal{G}}(k).$$

Given $\mathcal{G}^+$ (which is acyclic), we have $\theta = (\theta_{\mathcal{E}^+}, \theta_{\mathcal{I}^+})$, where $\theta_{\mathcal{E}^+} = (U_{\mathcal{E}^+}, W_{\mathcal{I}^+})$ are the (effective) parameters and $\theta_{\mathcal{I}^+} = (U_{\mathcal{I}^+}, W_{\mathcal{I}^+}) = 0$. Then the log-likelihood (4) becomes

$$L((\theta_{\mathcal{E}^+}, 0), \Sigma) = -\sum_{j=1}^p \sum_{i=1}^n \left( Y_{ij} - \sum_{(k, j) \in \mathcal{E}^+} U_{kj} Y_{ik} - \sum_{(l, j) \in \mathcal{I}^+} W_{lj} X_{il} \right)^2 / 2\sigma_j^2 + n \log(\sigma_j),$$

which involves $|\mathcal{E}^+| + |\mathcal{I}^+|$ parameters for $\theta_{\mathcal{E}^+}$. 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 $U$. 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

$$Y = V^T X + \epsilon, \quad \epsilon = (I - U^T)^{-1} \epsilon \sim N(0, \Omega^{-1}),$$

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

**Proposition 2** Suppose Assumption 1 is satisfied.

A. If $V_{ij} \neq 0$, then $X_j$ 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 $X_j$ such that $V_{ij} \neq 0$ and $V_{ij'} = 0$ for $j' \neq j$.

The proof of Proposition 2 is deferred to Appendix B.2. Intuitively, $V_{ij} \neq 0$ implies the dependence of $Y_j$ on $X_j$ through a directed path $X_j \rightarrow \cdots \rightarrow Y_j$ and hence that $X_j$ 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 $V$ and construction of $\mathcal{G}^+$.

#### 3.1.1 NODEWISE CONSTRAINED REGRESSIONS

We estimate $V = (V_1, \ldots, V_p)$ via nodewise $\ell_0$-constrained regressions,
\[ \hat{V}_j = \arg \min_{V_j} \sum_{i=1}^{n} (Y_i - V_j' X_i)^2 \quad \text{s.t.} \quad \sum_{i=1}^{q} I(V_{ij} \neq 0) \leq \kappa_j; \quad j = 1, \ldots, p, \]

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 \( I(z \neq 0) \) (Shen et al., 2012) and develop a difference-of-convex (DC) program with the \( \ell_0 \)-projection to improve the globality of the solution of (7). Specifically, at the \((t+1)\)th iteration, given \( \hat{V}_j^{[t]} \), we solve the weighted Lasso problem,

\[ \hat{V}_j^{[t+1]} = \arg \min_{V_j} \sum_{i=1}^{n} (Y_i - V_j' X_i)^2 + 2n \gamma_j \tau_j \sum_{l=1}^{q} I(\hat{V}_{lj}^{[t]} \neq 0) \leq \tau_j; \quad j = 1, \ldots, p, \]

where \( \gamma_j > 0 \) is an internal hyperparameter used by the DC program; see Remark 2 below. The DC program terminates at \( \hat{V}_j = \hat{V}_j^{[t]} \) such that \( \|\hat{V}_j^{[t+1]} - \hat{V}_j^{[t]}\|_\infty \leq \text{tol} \) or \( t \) achieves the maximum iteration number, where tol is the machine precision. Then, the solution \( \hat{V}_j \) of (7) is computed by projecting \( \hat{V}_j \) onto the set \( \{v \in \mathbb{R}^q: \|v\|_0 \leq \kappa_j\} \).

Algorithm 1 summarizes the computation method.

**Algorithm 1:** Constrained estimation via DC program + \( \ell_0 \) projection.

Input: data \( Y \) and \( X \).

Parameters: \( \{(\kappa_j, \tau_j)\}_{1 \leq j \leq s} \), candidate values \( \{\gamma(1) > \cdots > \gamma(R)\} \) for \( \gamma_j \).

Output: the estimate \( \hat{V} \).

1. for each \( 1 \leq j \leq p \) do
   2.    for each \( 1 \leq r \leq R \) do
      3.      Initialize \( \hat{V}_j^{(r)} = 0 \), and \( \gamma_j = \gamma(r) \);
      4.      Compute \( \hat{V}_j \) via DC program (8) with \( \hat{V}_j^{(r)} = 0 \);
      5.      Let \( B = \{i: \|\hat{V}_{ij}\| \neq 0 \} \) be the set of the largest \( \kappa_j \) elements of \( \|\hat{V}_{ij}\|_{0} \);
      6.      Compute \( \hat{V}_j^{(r)} = \arg \min_{v \in \mathbb{R}^q: \|v\|_0 = \kappa_j} \|Y_j - Xv\|_2^2 \);
   7. end
   8. Compute \( \hat{V}_j = \arg \min_{v \in \mathbb{R}^q: \|v\|_0 \leq \kappa_j} \|Y_j - Xv\|_2^2 \);
9. end
10. return \( \hat{V} = (\hat{V}_1, \ldots, \hat{V}_p) \).

**Remark 2** In Algorithm 1, \( \gamma_j \) is chosen from a set of candidate values \( \{\gamma(r)\}_{1 \leq s \leq R} \). In our implementation, \( \gamma_j \) is not directly tuned by the user and \( \{\gamma(r)\}_{1 \leq s \leq R} \) is provided by default. When \( \{(\kappa_j, \tau_j)\}_{1 \leq s \leq s} \) are suitably specified by the user, for any value \( \gamma(r) \) lies in the proper ranges, \( \{\hat{V}_1, \ldots, \hat{V}_p\} \) are the global solutions of (7) almost surely; see Theorem 4 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 \( V \).

Proposition 2 suggests that the leaf variables of \( \mathcal{G} \) (with their instruments) can be identified based on matrix \( V \). To proceed, let \( \mathcal{Z} \) and its complement \( \mathcal{Z}^c \) be (generic) nonempty subsets of \( \{1, \ldots, p\} \) such that \( Y_{\mathcal{Z}^c} \) are non-descendants of \( Y_{\mathcal{Z}} \). Define a sub-DAG \( \mathcal{G}_{\mathcal{Z}^c} = (Y_{\mathcal{Z}^c}, X; \mathcal{E}_{\mathcal{Z}^c}, \mathcal{I}_{\mathcal{Z}^c}) \), where \( \mathcal{E}_{\mathcal{Z}^c} \subseteq \mathcal{E} \) is the set of primary edges among \( Y_{\mathcal{Z}^c} \) and \( \mathcal{E}_{\mathcal{Z}^c} \subseteq \mathcal{I} \) is the set of intervention edges between \( X \) and \( Y_{\mathcal{Z}^c} \). The following proposition offers insights into the connection between \( V \) and \( \mathcal{G}_* \).
Proposition 3 Suppose Assumption 1 is satisfied. Let $Y_k$ be a leaf in $\mathcal{G}_\gamma$ and $Y_j$ be in $\mathcal{Y}_\gamma$.

A. If $\mathcal{V}_{ij} \neq 0$ for each instrument $X_l$ of $Y_k$ in $\mathcal{G}_\gamma$, we have $(k, j) \in \mathcal{E}_\gamma$.

B. If $Y_k$ is an unmediated parent of $Y_j$, then $\mathcal{V}_{ij} \neq 0$ for each instrument $X_l$ of $Y_k$ in $\mathcal{G}_\gamma$.

Proposition 3 (proved in Appendix B.3) together with Proposition 2 indicates that $\mathcal{G}_\gamma$ can be constructed from $\mathcal{V}$. 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 $\mathcal{G}_\gamma$ (where $\mathcal{Y}_\gamma$ are peeled variables). Then $\mathcal{E}_\gamma$ can be constructed by including all edges $(k, j)$ such that $Y_k$ is a leaf in the sub-DAG $\mathcal{G}_\gamma$, $Y_j$ is a peeled variable, and $\mathcal{V}_{ij} \neq 0$ for each instrument $X_l$ of leaf $Y_k$ in $\mathcal{G}_\gamma$. By Proposition 3, (A) confirms that all such edges are in $\mathcal{E}_\gamma$, 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, $\mathcal{E}_\gamma$ can be recovered from $\mathcal{V}$. Then $\mathcal{F}_\gamma$ is equal to $\{(l, j) : \mathcal{V}_{lk} \neq 0 \text{ if } k = j \text{ or } (k, j) \in \mathcal{E}_\gamma\}$.

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 $\mathcal{V}$

**Output:** estimated ARG $\hat{\mathcal{G}}_\gamma$

1. Initialize $\mathcal{V}_{\text{root}} \leftarrow \mathcal{V}$, $\hat{\mathcal{E}} \leftarrow \emptyset$, $\hat{\mathcal{X}} \leftarrow \{(i, l) : \mathcal{V}_{i l} \neq 0\}$.
2. Initialize $\mathcal{G}_{\text{root}}$ by $\mathcal{V}_{\text{root}} \leftarrow \{1, \ldots, q\}$, $\mathcal{X} \leftarrow \hat{\mathcal{X}}$, $\mathcal{E} \leftarrow \hat{\mathcal{E}}$.
3. While $\mathcal{V}_{\text{root}}$ is not empty do
   4. In $\mathcal{G}_{\text{root}}$, identify the instruments on leaves
      $\mathcal{B} \leftarrow \{i : \text{minimizes } \sum_{k \in \mathcal{A}} \mathcal{V}_{ik} \}$ and the leaves
      $\mathcal{E} \leftarrow \{k : \text{maximizes } \sum_{i \in \mathcal{B}} \mathcal{V}_{ik} \}$.
   5. Let $\mathcal{S}_k \leftarrow \{l : \mathcal{V}_{ik} \text{ maximizes } \mathcal{V}_{ik} \} / \{\mathcal{B} \}$ be the instruments for each $k \in \mathcal{E}$.
   6. Update $\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \{(i, j) : j \in \{1, \ldots, q\} \setminus \mathcal{X}, (k, j) \in \hat{\mathcal{E}} \}$.\(\text{and update } \mathcal{V}_{\text{root}} \text{ by keeping the columns in } \mathcal{V}_{\text{root}}\).
   7. Update $\hat{\mathcal{X}} \leftarrow \hat{\mathcal{X}} \cup \{(i, l) : (i, l) \notin \hat{\mathcal{X}}, (i, j) \in \hat{\mathcal{E}} \}$.
8. Return $\hat{\mathcal{E}}$ and $\hat{\mathcal{X}}$.

Remark 3 Given $\hat{\mathcal{G}}_\gamma = (\mathcal{Y}, \mathcal{X}; \hat{\mathcal{E}}_\gamma, \hat{\mathcal{F}}_\gamma)$, we estimate $(\mathcal{U}_{\hat{\mathcal{E}}_\gamma}, \mathcal{W}_{\hat{\mathcal{F}}_\gamma})$ column-wise from (5),

$$\min \sum_{i = 1}^{n} \left( Y_{ij} - \sum_{k \in \mathcal{A}} \mathcal{U}_{ij} Y_{ik} - \sum_{l \in \mathcal{B}} \mathcal{W}_{ij} X_{il} \right)^2 \text{ s.t. } \sum_{k \in \mathcal{A}} \mathcal{U}_{ik} \neq 0 \leq \kappa'_{ij}. \quad (9)$$

The final estimates are $\hat{\mathcal{U}} = (\mathcal{U}_{\hat{\mathcal{E}}_\gamma}, 0)$ and $\hat{\mathcal{W}} = (\mathcal{W}_{\hat{\mathcal{F}}_\gamma}, 0)$. In (9), the sparsity constraint is imposed to recover the nonzero elements of $\mathcal{U}$; 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 $\mathcal{G}_\gamma$. Next, we perform tests via data perturbation, accounting for the uncertainty of estimating $\mathcal{G}_\gamma$. 

}\[ J \text{ Mach Learn Res. Author manuscript; available in PMC 2023 September 12. }\]


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

$$\max_{\theta, \Sigma} \max_{G+} L(\theta, \Sigma).$$

Thus, we define the likelihood ratio by

$$L_r = L\left(\hat{\theta}^{(1)}, \hat{\Sigma}\right) - L\left(\hat{\theta}^{(0)}, \hat{\Sigma}\right) = \sum_{j=1}^p \left(\text{RSS}(\hat{\theta}^{(0)}_j) - \text{RSS}(\hat{\theta}^{(1)}_j)\right)/2\hat{\sigma}_j,$$

(10)

where $$\hat{\theta}^{(0)} = (\hat{\theta}^{(0)}_\mathcal{X}, 0)$$ and $$\hat{\theta}^{(1)} = (\hat{\theta}^{(1)}_\mathcal{X}, 0)$$ are MLEs (given $$\mathcal{M}$$) under $$H_0$$ and $$H_a$$ respectively, $$\mathcal{M} = (Y, X; \tilde{\mathcal{X}} \cup \mathcal{H}, \tilde{\mathcal{I}} \cup \mathcal{F}_+$$), is an estimate for $$G+$$, and $$\hat{\Sigma}$$ is an estimate for $$\Sigma$$. Instead of $$\mathcal{M} = (Y, X; \tilde{\mathcal{X}} \cup \mathcal{H}, \tilde{\mathcal{I}} \cup \mathcal{F}_+$$), the graph $$\mathcal{M}$$ (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, $$\mathcal{M}$$ may not be a DAG, and thus not all edges in $$\mathcal{H}$$ could present in the DAG parameterized by $$(\hat{\theta}^{(1)}_\mathcal{X}, 0)$$. As a result, $$L_r$$ 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.

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 $$\tilde{\mathcal{G}} = \mathcal{G}$$, and $$\tilde{\mathcal{M}} = (Y, X; \tilde{\mathcal{X}} \cup \mathcal{H}, \mathcal{F}_+)$$.

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

- $$H_0 : U_{45} = U_{53} = 0$$ versus $$H_a : U_{45} \neq 0$$ or $$U_{53} \neq 0$$, where $$\mathcal{H} = \{(4, 5), (5, 3)\}$$. In this case, $$\{(4, 5), (5, 3)\}$$ forms a cycle with the edges in $$\mathcal{X} \setminus \mathcal{H}$$, and $$\mathcal{M}$$ is cyclic. Given $$\mathcal{M}$$, the likelihood tends to be maximized under either DAG $$(Y, X; \tilde{\mathcal{X}} \cup \{(4, 5)\}, \mathcal{F}_+)$$ or DAG $$(Y, X; \tilde{\mathcal{X}} \cup \{(5, 3)\}, \mathcal{F}_+)$$ when data is sampled under $$H_0$$. Thus, we have

$$L(\hat{\theta}^{(1)}, \hat{\Sigma}) = \max\{L(U_{45} = U_{53} = 0, \tilde{\mathcal{X}}, \tilde{\Sigma}), L(U_{45} = 0, U_{53} = 0, \tilde{\mathcal{X}}, \tilde{\Sigma})\}.$$
As a result, the likelihood ratio distribution becomes complicated in this situation due to the dependence between the two components in $L(\theta^{(1)}, \Sigma)$.

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

**Definition 3 (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 4** In practice, $\mathcal{D}$ is unknown and needs to be estimated from data. Indeed, $\mathcal{D}$ can be computed based on the estimated ARG $\tilde{\mathcal{G}}$, because a directed edge $(k, j)$ is nondegenerate if and only if $\{(k, j)\} \cup \tilde{\mathcal{E}}$ 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 $\hat{\theta}^{(1)}$ is the MLE subject to the DAG $\hat{\mathcal{G}} = (Y, X; \tilde{\mathcal{E}}, \tilde{\mathcal{F}})$, and $\hat{\theta}^{(0)}$ is the MLE subject to an additional constraint $U_{\mathcal{R}} = 0$. The likelihood ratio (10) can be further simplified.

Let $D_S(j) = \{k: (k, j) \in \tilde{\mathcal{D}}\}$ in DAG $\hat{\mathcal{G}}$, where $\tilde{\mathcal{D}}$ is the estimated set of nondegenerate edges of $H_0$ with respect to $\mathcal{G}$. Furthermore, observe that if $D_S(j) = \emptyset$, then $RSS(\hat{\theta}^{(0)}) = RSS(\hat{\theta}^{(1)})$. Hence, $L_r$ only summarizes the contributions from the primary variables with the (estimated) nondegenerate hypothesized edges,

$$L_r = \sum_{\{j:D_S(j) \neq \emptyset\}} (RSS(\hat{\theta}^{(0)}) - RSS(\hat{\theta}^{(1)}))/2\hat{\sigma}_j^2,$$

where we estimate $\Sigma = \text{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$ by

$$\hat{\sigma}_j^2 = RSS(\hat{\theta}^{(1)})/(n - |PA_S(j)| - |IN_S(j)|), \quad j = 1, \ldots, p.$$

(11)
The likelihood ratio (11) for testing directed edges (2) requires an estimation of $\mathcal{G}_\pi$ (and $\pi'$), where we must account for the uncertainty of $\hat{\mathcal{G}}_\pi$ (and $\hat{\pi}'$) for finite-sample inference. To proceed, we consider the test statistic $L_r$ based on a “correct” ARG $\mathcal{M} \supseteq \mathcal{G}_\pi$, where $\mathcal{M} = (Y, X; \mathcal{A}, \mathcal{E}) \supseteq G = (Y, X; \mathcal{E}_\pi, \mathcal{J}, \mathcal{C})$ means that $\mathcal{A} \supseteq \mathcal{E}_\pi$, and $\mathcal{C} \supseteq \mathcal{J}$. 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 $\mathcal{G}_\pi$. By comparison, a “wrong” ARG $\mathcal{M} \not\supseteq \mathcal{G}_\pi$ provides an incorrect topological order, and a test based on a “wrong” ARG may be biased, accompanied by an inflated type-I error.

Let $\mathbf{Z} = (Y, X)$ denote the data matrix of primary and intervention variables and let $\mathbf{e} = (e_1, \ldots, e_n)^T$ denote the error matrix, where the rows $\{\mathbf{z}_i = (Y_i, X_i)\}_{1 \leq i \leq n}$ and $\{e_i\}_{1 \leq i \leq n}$ are sampled independently from (1). From (11), assuming $\hat{\mathcal{G}}_\pi \supseteq \mathcal{G}_\pi$ is a “correct” ARG, the likelihood ratio becomes

$$L_r = \sum_{\{j: D_\pi(j) \neq \emptyset\}} \frac{\| (P_{\hat{A}_j} - P_{\hat{A}_j})^{1/2} Y \|^2}{2\| (I - P_{\hat{A}_j})^{1/2} Y \|^2 (n - |\hat{A}_j|)} \sum_{\{j: D_\pi(j) \neq \emptyset\}} \frac{\| (P_{\hat{A}_j} - P_{\hat{A}_j})^{1/2} e \|^2}{2\| (I - P_{\hat{A}_j})^{1/2} e \|^2 (n - |\hat{A}_j|)},$$

where $P_{\hat{A}_j}$ is the projection matrix onto the column span of $\mathbf{Z}_{-\mathbf{j}}$, $\hat{A}_j = PA_\pi(j) \cup IN_\pi(j)$, and $\tilde{B}_j = (PA_\pi(j) \cup IN_\pi(j)) \setminus D_\pi(j)$ for $1 \leq j \leq p$. In (13), we have $Y_{-j,0:0,j} = 0$ for all $j$ under the null hypothesis $H_0$, while $Y_{-j,0:0,j} \neq 0$ for some $j$ under the alternative hypothesis $H_\pi$, and thus $L_r$ tends to be large under $H_\pi$.

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

$$L_r^* = \sum_{\{j: D_\pi(j) \neq \emptyset\}} \frac{\| (P_{\hat{A}_j} - P_{\hat{A}_j})^{1/2} e' \|^2}{2\| (I - P_{\hat{A}_j})^{1/2} e' \|^2 (n - |\hat{A}_j|)}.$$

Note that (14) mimics (13) when $Y_{-j,0:0,j} = 0$. As a result, when $\hat{\mathcal{G}}_\pi \supseteq \mathcal{G}_\pi$, the conditional distribution of $L_r^*$ given the data $\mathbf{Z}$ well approximates the null distribution of $L_r$, where
the model selection effect is accounted for by assessing the variability of \((\tilde{\mathcal{A}}, \tilde{\mathcal{B}})_{1 \leq s \leq p}\) over different realizations of \((Z^\ast, e^\ast)\).

In practice, we use Monte-Carlo to approximate the distribution of \(Lr^\ast\) given \(Z\). We generate \(M\) perturbed samples \((\{Z_s, e_s\}_{1 \leq s \leq M})\) independently and compute \((Lr^\ast_{1 \leq s \leq M})\), respectively. Then, we examine the condition \(\mathcal{E}^\ast_{s, a} \supseteq \mathcal{E}_a\) by checking its empirical counterpart \(\mathcal{E}^\ast_{s, a} \supseteq \mathcal{E}_a\). The DP p-value of the edge test in (2) is defined as

\[
Pval = \left( \sum_{m=1}^{M} I(Lr^\ast_{m} \geq Lr, \mathcal{E}^\ast_{s, a} \supseteq \mathcal{E}_a) \right) / \left( \sum_{m=1}^{M} I(\mathcal{E}^\ast_{s, a} \supseteq \mathcal{E}_a) \right),
\]

where \(I(\cdot)\) is the indicator function.

**Remark 5** Instead of (14), a naive approach is to recompute the likelihood ratio by treating the perturbed sample \(Z^\ast\) as \(Z\) while not using the information of \(e^\ast\). However, this is infeasible. For explanation, assuming \(\mathcal{E}^\ast_{s, a} \supseteq \mathcal{E}_a\) is a “correct” ARG, then this naive likelihood ratio is equal to

\[
\sum_{\{j : D^\ast_j \not\in \emptyset\}} \frac{\left\| (P^\ast_{j} - P^\ast_{\emptyset})^{1/2}(Y_j + e^\ast_j) \right\|^2}{2\left\| (I - P^\ast_j)^{1/2}(Y_j + e^\ast_j) \right\|^2/(n - |A_j|)}.
\]

Note that \(\{Y_{j}\}_{1 \leq s \leq p}\) given \(Z\) are deterministic and do not vanish under either \(H_0\) or \(H_a\).

Thus, the conditional distribution of this naive likelihood ratio given \(Z\) 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 \(\mathcal{H} = \{(k_1, j_1), \ldots, (k_m, j_m)\}\). Then the test of pathways in (3) can be reduced to testing sub-hypotheses

\[
H_{0, v} : U_{k_{v}, j_{v}} = 0, \quad \text{versus} \quad H_{a, v} : U_{k_{v}, j_{v}} \neq 0, \quad v = 1, \ldots, |\mathcal{H}|,
\]

where testing each sub-hypothesis is a directed edge test. Given \((\hat{\theta}, \hat{\Sigma})\), the likelihood ratio for \(H_{0, v}\) is \(Lr_v = L(\hat{\theta}^{(1)}, \hat{\Sigma}^\ast) - L(\hat{\theta}^{(0, v)}, \hat{\Sigma})\), where \(\hat{\theta}^{(0, v)}\) is the MLE under the constraint that \(U_{k_{v}, j_{v}} = 0\). When \(\mathcal{E}^\ast_{s, a} \supseteq \mathcal{E}_a\), we have

\[
Lr_v = \frac{\left\| (P^\ast_{k_v} - P^\ast_{\emptyset})^{1/2}(Y_{k_v} + e_{k_v}) \right\|^2}{2\left\| (I - P^\ast_{k_v})^{1/2}e_{k_v} \right\|^2/(n - |A_{k_v}|)}, \quad Lr^\ast_v = \frac{\left\| (P^\ast_{k_v} - P^\ast_{\emptyset})^{1/2}e^\ast_{k_v} \right\|^2}{2\left\| (I - P^\ast_{k_v})^{1/2}e^\ast_{k_v} \right\|^2/(n - |A_{k_v}|)},
\]

where the distributions of \(Lr_v^\ast\) given \(Z\) approximates the null distributions of \(Lr_v\). Finally, define the p-value of the pathway test in (3) as
Pval = \max_{1 \leq r \leq \ell} \left( \frac{\sum_{m = 1}^{M} I(Lr_{\nu,m} \geq L_{\nu}, \hat{S}_{\nu,m} \supseteq \hat{S})}{\sum_{i = 1}^{M} I(S_{m} \supseteq S)} \right). \tag{17}

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

Algorithm 3 summarizes the DP method for hypothesis testing.

| Algorithm 3: DP likelihood ratio test |
|--------------------------------------|
| **Input:** hypothesis $H_0$, data $Z$ |
| **Parameters:** Monte Carlo size $M$; parameters for Algorithm 1. |
| **Output:** the pivotal $Pval$. |
| 1. Compute $(\hat{S}, \ldots, \hat{S})$ and $L_r$ with data $Z$. |
| 2. For each $1 \leq m \leq M$ in parallel do |
| 3. Generate perturbed data $(Z^*_m, \epsilon^*_m)$; |
| 4. For (2), compute $L^*_m$ based on (16); |
| 5. For (3), compute $[L^*_m]_{1 \leq j \leq q}$ based on (16); |
| 6. end |
| 7. Compute $Pval$ as (15) or (17) accordingly; |
| 8. return $Pval$. |

**Remark 6** For acceleration, we parallelize Step 2 in Algorithm 3. Additionally, we use the estimate $\hat{\theta}$ as a warm-start initialization for the DP estimates, effectively reducing the computing time.

**Remark 7 (Connection with bootstrap)** One may consider parametric or nonparametric bootstrap for $L_r$. The parametric bootstrap requires a good initial estimate of $(U, W)$. 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 $L_r$, while DP provides a better approximation of $L_r$, 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 $\zeta$ be a generic vector and $\zeta_A$ be the subvector of $\zeta$ with coordinates in $A$. Let $\zeta' = \|V\|_2$, and $\kappa_{\max} = \max_{1 \leq j \leq p} \kappa_j$.

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

A. \[
\min_{\{A: |A| \leq 2 \kappa_{\max}\}} \min_{\{\zeta: \|\zeta\| \leq 3 \|\zeta\|\}} \frac{\|X\zeta\|_2^2 / n \|\zeta\|_2}{\zeta'} \geq c_1. \]

B. \[
\max_{1 \leq i \leq q} \left( \frac{X^TX}{n} \right)^{\frac{1}{2}} / n \leq c_2. \]

**Assumption 3** \[
\min_{V_i \neq 0} \left[ V_i \|100c_1^{-1}c_2 \max_{1 \leq j \leq p} (\Omega^{-1/2})_j \log(q) / n + \log(n) / \sqrt{n} \right]. \]

J Mach Learn Res. Author manuscript; available in PMC 2023 September 12.
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 is satisfied with a probability tending to one for isotropic subgaussian or bounded $X$ (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 4 (Finite termination and consistency) Suppose Assumptions 1-3 are met with constants $c_1 < 6c_2$, and the machine precision $\text{tol} \ll 1/n$ is negligible. For $1 \leq j \leq p$, if the tuning parameters $(\kappa_j, \tau_j)$ of Algorithm 1 are suitably chosen such that

$$
\kappa_j = \kappa_j', \frac{36c_2}{c_1} \sqrt{\frac{\log(q)}{n}} \leq \tau_j \leq \frac{2}{\min_{V_{ij} \neq 0} |V_{ij}|^{1/2}}
$$

then for any $\gamma_j$ such that $\gamma_j (32c_2^2 \Omega_{jj}^{-1} n^{-1} (\log(q) + \log(n)))^{1/2} \leq \gamma_j \leq c_2^6$, almost surely we have Algorithm 1 yields a global solution $\hat{V}$, of (7) in at most $1 + \lceil \log(\kappa_{\max})/\log(16) \rceil$ 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 $\mathcal{J}^+$ when $n$ is sufficiently large.

In view of Theorem 4 (proved in Appendix B.4), it suffices to specify the maximum number of DC iterations as $1 + \lceil \log(\kappa_{\max})/\log(4) \rceil$. Then the time complexity of Algorithm 1 is $p \times O(\log(\kappa_{\max}) \times O(q^3 + nq^2)$, 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 $O(p \times \log(\kappa_{\max}) \times (q^3 + nq^2))$. Finally, the peeling method does not apply to observational data ($W = 0$). In a sense, interventions are essential.

Theorem 4 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 $\delta = (Y, X; \mathcal{E}, \mathcal{D}, \mathcal{F}_i)$.

**Assumption 4**

$$
\max_{\{j : D_j \neq \emptyset\}} (|\text{PA}_i(j)| + |\text{IN}_i(j)|)/n \leq \rho \text{ as } n \to \infty, \text{ for a constant } 0 < \rho < 1.
$$

Assumption 4 is a hypothesis-specific condition restricting the underlying dimension of the testing problem. Usually, $|\text{PA}_i(j)| = |\text{AN}_i(j)| = |\text{IN}_i(j)| = \kappa_j \ll p$, $1 \leq j \leq p$, which relaxes the condition $n \gg \log(p)/\sqrt{|\mathcal{E}|}$ for the constrained likelihood ratio test (Li et al., 2020).

**Theorem 5 (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 4.
A. For the test of directed edges (2),

\[
\lim_{n \to \infty} \lim_{M \to \infty} \mathbb{P}(P_{\theta} < \alpha) = a, \text{ if } H_0 \text{ is nondegenerate.}
\]

\[
\lim_{n \to \infty} \lim_{M \to \infty} \mathbb{P}(P_{\theta} = 1) = 1, \text{ if } H_0 \text{ is degenerate.}
\]

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

\[
\limsup_{n \to \infty} \lim_{M \to \infty} \mathbb{P}(P_{\theta} < \alpha) = a, \text{ if } H_0 \text{ is nondegenerate with } |\mathcal{D}| = |\mathcal{H}|.
\]

\[
\limsup_{n \to \infty} \lim_{M \to \infty} \mathbb{P}(P_{\theta} = 1) = 1, \text{ if } |\mathcal{D}| < |\mathcal{H}|.
\]

By Theorem 5 (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 6 (proved in Appendix B.6) summarizes the asymptotics for directed edge test (2).

**Proposition 6 (Asymptotics of edge test)** Suppose the assumptions of Theorem 5 are met. Under a nondegenerate and regular $H_0$, as $n \to \infty$,

A. \[2Lr \xrightarrow{\text{d}} \chi^2_m, \text{ if } |\mathcal{D}| > 0 \text{ is fixed.}\]

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

**Remark 8** As opposed to the entire $\mathcal{G}$ (or $\mathcal{D}$), Theorem 5 and Proposition 6 only require correct identification of the local structures $\{\text{AN}(\varnothing), \text{IN}(\varnothing), \text{DS}(\varnothing)\}$: $\mathcal{D} \neq \emptyset$. 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 $\theta^* = (U^*, W^*)$ satisfies $H_0$. Let $\Delta \in \mathbb{R}^{p \times p}$ satisfy $\Delta_{\varnothing} = 0$ so that $U^* + \Delta$ represents a DAG.

For nondegenerate and regular $H_0$, consider an alternative $H_a: U_{\varnothing} = U_{\varnothing} + \Delta_{\varnothing}$, and define the power function as

\[
\beta(\theta^*, \Delta) = \mathbb{P}_{H_a}(P_{\theta} < \alpha).
\]

**Proposition 7 (Local power of edge test)** Suppose $H_0$ is nondegenerate and regular. Let $\|\Delta\|_F = \|\Delta_{\varnothing}\|_F = n^{-1/2}\delta$ when $|\mathcal{D}| > 0$ is fixed and $\|\Delta\|_F = |\mathcal{D}|^{1/4} n^{-1/2} h$ when $|\mathcal{D}| \to \infty$, where $\delta > 0$ and $\|\cdot\|_F$ is the matrix Frobenius norm. If the assumptions of Theorem 5 are met, then under $H_a$, as $n, M \to \infty$,

\[
\beta(\theta^*, \Delta) \geq \begin{cases} 
\mathbb{P}_{Z \sim N(0, I)}[\|Z + \sqrt{n}\Delta\|_2^2 > \chi^2_{1/2 - \delta}] & \text{if } |\mathcal{D}| > 0 \text{ is fixed,} \\
\mathbb{P}_{Z \sim N(0, I)}[Z > z_{1/2 - \delta} \|\Delta\|_F / \sqrt{|\mathcal{D}|}] & \text{if } |\mathcal{D}| \to \infty, \frac{\|\Delta\|_F \log|\mathcal{D}|}{n} \to 0.
\end{cases}
\]

J Mach Learn Res. Author manuscript; available in PMC 2023 September 12.
where $\chi_{\frac{1}{2}, 1 - \alpha}$ and $z_{1 - \alpha}$ denote the $(1 - \alpha)$th quantile of distributions $\chi_2$ and $N(0, 1)$, respectively. Hence, $\lim_{\delta \to \infty} \lim_{n \to \infty} \beta(\theta^*, \Delta) = 1$.

**Proposition 8 (Local power of pathway test)** Suppose $H_0$ is nondegenerate and regular with $|\mathcal{D}| = |\mathcal{H}|$. Let $\min_{k, j < p} |U_{kj} + \Delta_{kj}| = n^{-1/2} \delta$ when $|\mathcal{H}| > 0$ is fixed and $\min_{k, j < p} |U_{kj} + \Delta_{kj}| = n^{-1/2} \sqrt{\log |\mathcal{H}|}$ when $|\mathcal{H}| \to \infty$. If the assumptions of Theorem 5 are met, then under $H_a$, as $n, M \to \infty$,

$$
\beta(\theta^*, \Delta) \geq 1 - \frac{|\mathcal{H}|}{\sqrt{2\pi}} \exp \left( - \frac{1}{2} \left[ \delta \sqrt{\log |\mathcal{H}|} \max_{1 \leq j \leq p} \Omega_j - \chi_{\frac{1}{2}, 1 - \alpha}^2 \right] \right),
$$

where $\chi_{\frac{1}{2}, 1 - \alpha}$ is the $(1 - \alpha)$th quantile of distribution $\chi_2$. Hence, $\lim_{\delta \to \infty} \lim_{n \to \infty} \beta(\theta^*, \Delta) = 1$.

The proofs of Propositions 7 and 8 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 $U \in \mathbb{R}^{p \times p}$, 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 $U_{1,2j+1} = 1$ and $U_{2,2j+2} = 1$ 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 $W \in \mathbb{R}^{q \times p}$, 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 $W = (A^T, B^T, 0^T)^T$, where $A, B \in \mathbb{R}^{p \times p}$ and $0 \in \mathbb{R}^{(q - 2p) \times p}$.

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

- **Setup (B).** Set $A_{j,j+1} = B_{j,j+1} = B_{j,j+1} = 1$; $j = 1, \ldots, p - 1$, $A_{pp} = 1$, while other entries of $A, B$ 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 \((Y, X)\) for each setup, we sample \(X \sim N(0, \Sigma_X)\) with \(\Sigma_X = 0.5^{l-l'} I; 1 \leq l, l' \leq q\) and sample \(Y\) according to (1) with \((U, W, \sigma_1^2, ..., \sigma_p^2)\), where \(\sigma_1^2, ..., \sigma_p^2\) 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 \(L_r\), while OLR uses \(L_r(S, \Sigma)\) assuming that \(S\) were known in advance. The p-values of LR and OLR are computed via Proposition 6. 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:
  
  \[
  \begin{align*}
  &i. \quad H_0: U_{1,20} = 0 \text{ versus } H_a: U_{1,20} \neq 0. \text{ In this case, } |D| = 1. \\
  &ii. \quad H_0: U_x = 0 \text{ versus } H_a: U_x \neq 0, \text{ where } X = \{(k, 20): k = 1, ..., 15\}. \text{ In this case, } |D| = 15.
  \end{align*}
  \]

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

- **Test of directed pathways.** We test the directed path \(Y_1 \rightarrow Y_5 \rightarrow Y_{10} \rightarrow Y_{15} \rightarrow Y_{20}\), namely \(X = \{(1, 5), (5, 10), (10, 15), (15, 20)\}\) in (3). Since (3) is a test of composite null hypothesis, the data are generated under a graph with parameters \((U, W, \Sigma)\) satisfying \(H_0\), where \(U_x = 0\). Five hypotheses \(H_a: U_x = 0.1^l; l = 1, 2, 3, 4, 5\), 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 \(\tilde{S}_+ \supseteq S_+\), which overestimates \(S_+\), 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.

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 \( U \). 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 \( W = (I_p, 0)^\top \in \mathbb{R}^{q \times p} \). 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)

\[
\text{SHD}(\hat{U}, U) = \sum_{k,j} |I(\hat{U}_{kj} \neq 0) - I(U_{kj} \neq 0)|,
\]

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

\[
H_0: U_{1,20} = 0 \text{ versus } H_a: U_{1,20} = 1/\sqrt{n}.
\]

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 \( \hat{U}_{1,20} \neq 0 \).

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 10, suggesting that consistent reconstruction requires the smallest size of nonzero coefficients to be of order $\sqrt{\log(n)/n}$ with the tuning parameter $\tau$ of the same order (fixing $p, q$). In this case, the edge $U_{1,20}$ is of order $1/\sqrt{n}$, which is less likely to be reconstructed as $n$ increases. In contrast, Proposition 7 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 comparison, 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; Candès 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, in 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 \( H_0 : U_{kj} = 0 \) versus \( H_a : U_{kj} \neq 0 \), for each edge \((k, j)\) as shown in Figure 6 (a) and (b). Moreover, we consider two hypothesis tests of pathways \( H_a : U_{kj} = 0 \) for some \((k, j) \in \mathcal{P}_\ell\) versus \( H_a : U_{kj} \neq 0 \) for all \((k, j) \in \mathcal{P}_\ell; \ell = 1, 2\), where the two pathways are specified by \( \mathcal{P}_1 = \{ \text{PSEN1} \rightarrow \text{CAPN1} \rightarrow \text{CDK5R1} \} \), and \( \mathcal{P}_2 = \{ \text{PSEN1} \rightarrow \text{CAPN2} \rightarrow \text{CDK5R1} \} \). See Figure 7. Of note, for clear visualization, Figure 6 (a), (b) and Figure 7 only display the edges related to hypothesis testing, whereas 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 \rightarrow APBB1, APP \rightarrow GSK3B, FADD \rightarrow CASP3\} in the AD-MCI group, but no evidence in the CN group. Meanwhile, this test suggests the presence of connections \{TNFRSF1A \rightarrow CASP3, FADD \rightarrow CASP8\} in the CN group but not so in the AD-MCI group. In both groups, we identify directed connections \{TNFRSF1A \rightarrow FADD, TNFRSF1A \rightarrow CASP8\}. In Figure 7, the pathway test (3) supports the presence of a pathway PSEN1 \rightarrow CAPN1 \rightarrow 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 \rightarrow CAPN2 \rightarrow 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 \( \text{APP} \rightarrow \text{GSK3B} \) in 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 \( \text{APP} \rightarrow \text{APBB1} \) also differs in AD-MCI and CN groups, which appears consistent with Figure 3 of Bu (2009). Moreover, the connection \( \text{CAPN1} \rightarrow \text{CDK5R1}, \) in the pathway \( \text{PSEN1} \rightarrow \text{CAPN1} \rightarrow \text{CDK5R1} \) 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 → CAPN1 → 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 8) 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 the 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
\[(U, W, \Sigma) : U \in \mathbb{R}^{p \times p} \text{ represents a DAG, } W \in \mathbb{R}^{p \times q}, \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2) \}.

As suggested by Proposition 1, 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

\[ Y = V^T X + \epsilon_V, \quad \epsilon_V = (I - U^T)^{-1} \epsilon \sim N(0, \Omega^{-1}), \]

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

**Example 2 (Identifiability)** In model (1), consider two non-identifiable bivariate situations:

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

| Model       | \( Y_1 \) | \( Y_2 \) |
|-------------|-----------|-----------|
| \( \theta \) | \( Y_1 = X_1 + X_2 + X_3 + \epsilon_1, \) | \( Y_2 = Y_1 - X_2 + X_1 + X_4 + \epsilon_3, \) |
| \( \tilde{\theta} \) | \( Y_1 = 0.5Y_2 + 0.5X_1 + X_2 - 0.5X_1 + \tilde{\epsilon}_1, \) | \( Y_2 = X_1 + 2X_3 + X_4 + \tilde{\epsilon}_2, \) |

where \( \epsilon_1, \epsilon_2 \sim N(0, 1) \) are independent, and \( \tilde{\epsilon}_1 \sim N(0, 0.5), \tilde{\epsilon}_2 \sim N(0, 2) \) are independent. As depicted in Figure 9, (19) satisfies Assumption 1C. However, Assumption 1B is violated given that \( \text{Cov}(Y_2, X_2 | X_{(1,3,4)}) = 0 \) 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 \( \text{Cov}(Y_1, X_4 | X_{(1,2,3)}) = 0 \) violating Assumption 1B. In this case, it can be verified that \( \theta \) and \( \tilde{\theta} \) correspond to the same distribution \( \mathbb{P}(Y | X) \), because they share the same \( (V, \Omega) \) even with different values of \( (U, W, \Sigma) \). 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:

| Model       | \( Y_1 \) | \( Y_2 \) |
|-------------|-----------|-----------|
| \( \theta \) | \( Y_1 = X_1 + X_2 + \epsilon_1, \) | \( Y_2 = Y_1 + X_2 + \epsilon_3, \) |
| \( \tilde{\theta} \) | \( Y_1 = 0.5Y_2 + 0.5X_1 + \tilde{\epsilon}_1, \) | \( Y_2 = X_1 + 2X_3 + \tilde{\epsilon}_2, \) |

where \( \epsilon_1, \epsilon_2 \sim N(0, 1) \) are independent, and \( \tilde{\epsilon}_1 \sim N(0, 0.5), \tilde{\epsilon}_2 \sim N(0, 2) \) 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, \( \theta \) and \( \tilde{\theta} \) yield the same distribution \( \mathbb{P}(Y | X) \) because they share the same \( (V, \Omega) \) even with different

---

Li et al. Page 23

J Mach Learn Res. Author manuscript; available in PMC 2023 September 12.
values of $(U, W, \Sigma)$. 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 $p = q = 5$,

$$
Y_1 = X_1 + \varepsilon_1, \quad Y_2 = 0.5Y_1 + X_3 + \varepsilon_2, \quad Y_5 = X_4 + \varepsilon_5, \\
Y_3 = 0.5Y_2 + X_5 + \varepsilon_3, \quad Y_4 = 0.5Y_3 - 0.1Y_1 + X_2 + \varepsilon_4,
$$

(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 $\hat{V}$ by Algorithm 1.

In particular,

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

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 \( Y_4 \) (\( X_2 \rightarrow Y_4 \)) because \( \hat{V}_{24} \neq 0 \) is the only nonzero in the row 2 with the smallest (positive) row \( \ell_0 \)-norm.

  – \( X_4 \) is identified as an instrument of leaf node \( Y_5 \) (\( X_4 \rightarrow Y_5 \)) because \( \hat{V}_{45} \neq 0 \) is the only nonzero in row 4 with the smallest (positive) row \( \ell_0 \)-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 \( Y_3 \) (\( X_5 \rightarrow Y_3 \)) in \( \mathcal{G}^{\text{work}} \) given that \( \hat{V}_{53} \neq 0 \) is the only nonzero element in the row with the smallest (positive) row \( \ell_0 \)-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 \( \hat{V}_{53} \neq 0 \), we have \( (3, 4) \in \mathcal{E}_+ \) by Proposition 3. 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 \( Y_2 \) (\( X_3 \rightarrow Y_2 \)) similarly in \( \mathcal{G}^{\text{work}} \) given that \( \hat{V}_{32} \neq 0 \) 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 \( \hat{V}_{32} \neq 0 \), we have \( (2, 3) \in \mathcal{E}_+ \). 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 \( Y_1 \) (\( X_1 \rightarrow Y_1 \)).

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 \( \hat{V}_{13} \neq 0 \), we have \( (1, 2) \in \mathcal{E}_+ \). Then \( \{ Y_1 \} \) is removed, and the peeling process is terminated.

Finally, Steps 9 and 10 identify

\[
\mathcal{E}_+ = \{(1, 2), (2, 3), (3, 4), (1, 3), (1, 4), (2, 4)\},
\]

\[
\mathcal{F}_+ = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 4), (3, 2), (3, 3), (3, 4), (4, 5), (5, 3), (5, 4)\},
\]

which are equal to \( \mathcal{E}_+ \) and \( \mathcal{F}_+ \), respectively.
In Example 3 \( \{(l,j) : V_{lj} \neq 0\} \neq \{(l,j) : V_{lj} \neq 0\} \), suggesting that the selection consistency of \( \widehat{V} \) is unnecessary for Algorithm 2 to correctly reconstruct the ARG \( \mathcal{G}_x \); see also Section A.3 for the theoretical justification.

### A.3 Relaxation of Assumption 3

Assumption 3 in Theorem 4 leads to consistent identification for \( V \). Now, we discuss when Algorithm 2 correctly reconstructs \( G^+ \) without requiring Assumption 3.

#### Assumption 5

For \( 1 \leq j \leq p \), there exists \( \tau_j^* \) such that

A. \( \{l : X_l \text{ intervenes on } Y_j \text{ or its unmediated parents} \} \subseteq \{l : |V_{lj}| \geq \tau_j^*\} \).

B. \( \tau_j^* \geq 100c_1^{-1} c_2(\Omega_{jj}^{-1/2})\sqrt{(|\{l : |V_{lj}| \geq \tau_j^*\}| + 1)(\log(q)/n + \log(n)/n)} \).

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

#### Theorem 9

Suppose Assumptions 1-2 and 5 are met with constants \( c_1 < 6c_2 \), and the machine precision \( \text{tol} \ll 1/n \) is negligible. For \( 1 \leq j \leq p \), there exist some suitable choice of tuning parameters \( (\kappa_j, \tau_j) \) in Algorithm 1 such that

\[
|\{l : |V_{lj}| \geq \tau_j^*\}| \leq \kappa_j \leq \kappa_j^* \leq \frac{36c_2}{c_1} \left( \Omega_{jj}^{1/2} \left( \frac{\log(q)}{n} + \frac{\log(n)}{n} \right) \right)^{1/2} \leq \frac{2\tau_j^*}{5},
\]

then for any \( \gamma_j \) such that

\[
\tau_j^* (32c_2^2 \Omega_{jj}^{-1} (\log(q) + \log(n)))^{1/2} \leq \gamma_j \leq c_1/6 \sqrt{\kappa_j^*} - |\{l : |V_{lj}| \geq \tau_j^*\}| + 1,
\]

almost surely we have Algorithm 1 terminates in at most \( 1 + \lceil \log(\kappa_{\text{max}}^*)/\log(4) \rceil \) 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 9 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 \( Z_i \rightarrow Z_j \) or \( Z_j \rightarrow Z_i \). A path (undirected) between \( Z_i \) and \( Z_j \) in \( \mathcal{G} \) is a sequence of distinct nodes \( (Z_k, \ldots, Z_j) \) such that all pairs of successive nodes in the sequence are adjacent. A nonendpoint node \( Z_k \) on a path \( (Z_i, \ldots, Z_k, Z_{k+1}, \ldots, Z_j) \) is a collider if \( Z_{k-1} \rightarrow Z_k \leftarrow Z_{k+1} \).
Otherwise it is a noncollider. Let $A \subseteq \{1, \ldots, p + q\}$, where $A$ does not contain $i$ and $j$. Then $Z_A$ 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 $Z_A$; or (ii) the path contains a collider that is not in $Z_A$ and has no descendant in $Z_A$. A node $Z_i$ is d-separated from $Z_j$ given $Z_A$ if $Z_A$ 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 $\varsigma$-strong faithful to a DAG with node set $\mathcal{V} = \{1, \ldots, p + q\}$ if

$$
\min_{A \subseteq \mathcal{V} \setminus \{i, j\}} \{|\text{Corr}(Z_i, Z_j | Z_A)|: Z_i \text{ is not d-separated from } Z_j \text{ given } Z_A\} > \varsigma,
$$

for $1 \leq i \neq j \leq p + q$, 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 $\varsigma \geq \sqrt{s_0 \log(p + q)/n}$, 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 $2^{(p+q-2)}$. Indeed, (24) is very restrictive in high-dimensional situations (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 the 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 $X \sim N(0, I)$. Consider model (1) with $p = q = 3$,

$$
Y_1 = W_{11} X_1 + \epsilon_1, \quad Y_2 = U_{12} Y_1 + W_{21} X_2 + \epsilon_2, \quad Y_3 = U_{13} Y_1 + U_{23} Y_2 + W_{33} X_3 + \epsilon_3,
$$

where $\epsilon_1, \epsilon_2, \epsilon_3 \sim N(0, 1)$ are independent and $U_{12}, U_{13}, U_{23}, W_{11}, W_{22}, W_{33} \neq 0$. Denote $Z = (Y_1, Y_2, Y_3, X_1, X_2, X_3)^\top$. Since the directed relations among $X$ are not of interest, (24) becomes

$$
\min_{Z_A = (Y_{A_1}, X_{A_2})} \{|\text{Corr}(Y_n, Y_j | Z_A)|: Y_n, Y_j \text{ are not d-separated given } Z_A\} > \varsigma,
$$

(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 $V_{11}, V_{12}, V_{13}, V_{22}, V_{23}, V_{33} > \sqrt{\log(q)/n}$, which in turn requires the minimal absolute value of six correlations $\geq \sqrt{\log(q)/n}$. 

---

*J Mach Learn Res.* Author manuscript; available in PMC 2023 September 12.
A.5 Irregular hypothesis

Assume, without loss of generality, that \( \widehat{\mathcal{D}} = \mathcal{D} \) and \( \widehat{\mathcal{E}}_+ = \mathcal{E} \) 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 \( H_0^{(1)}, \ldots, H_0^{(q)} \), each of which is regular. Then testing \( H_0 \) is equivalent to multiple testing for \( H_0^{(1)}, \ldots, H_0^{(q)} \).
  
  For instance, in Example 1, \( H_0 : U_{45} = U_{53} = 0 \) is irregular, and \( H_0 \) can be decomposed into \( H_0^{(1)} : U_{45} = 0 \) and \( H_0^{(2)} : U_{53} = 0 \).

- 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 \( \mathbf{Y}_{.,j} \) as the response variable and the input \( (\mathbf{Y}_{.,\text{AN}_+}, \mathbf{X}_{.,\text{ING}_+ \cup \mathcal{G}}) \) as the covariates for \( 1 \leq j \leq p \). The tuning parameters for solving (9) by Algorithm 1 are denoted by \( \{ (\kappa_j, \tau_j) \}_{\ell \in \mathcal{P}} \).

Let \( \bar{\kappa} = \max_{1 \leq j \leq p} |\text{AN}_{\kappa,j}| + |\text{N}_{\kappa,j}| \) and \( \mathbf{Z} = (\mathbf{Y}, \mathbf{X}) \).

**Assumption 6** For constants \( c_3, c_4 > 0 \),

\[
\begin{align*}
\text{A.} & \quad \min_{\{A: |A| \leq 2\bar{\kappa}\}} \min_{\{\xi: \|\xi\|_1 \leq 3\|\xi_0\|_1\}} \|\mathbf{Z}_{\xi:}\|/n\|\xi\|_1^2 \geq c_3, \\
\text{B.} & \quad \max_{1 \leq k \leq p + q} n^{-1}(\mathbf{Z}^T \mathbf{Z})_{kk} \leq c_4.
\end{align*}
\]

**Assumption 7** For \( |U_{ij}| \geq 100c_1^{-1}c_4 \), \( \max_{1 \leq j \leq p} (\sigma_j n^{1/2} \log(p)/n + \log(n)/n) \).

**Theorem 10** Suppose the assumptions in Theorem 4 are satisfied. In addition, suppose Assumptions 6–7 are met with constants \( c_3 < 6c_4 \). For \( 1 \leq j \leq p \), assuming \( |\text{AN}_{\kappa,j}| \ll n \) and \( |\text{N}_{\kappa,j}| \ll n \), if the tuning parameters \( (\kappa_j, \tau_j) \) are suitably chosen such that

\[
\kappa_j = |PA_{\kappa,j}|, \quad \frac{36c_4}{c_3} \sigma_j \sqrt{\frac{\log(p)}{n} + \frac{\log(n)}{n}} \leq \tau_j \leq \frac{2}{\Sigma \mathcal{U}_{ij} \neq 0} \min \mathcal{U}_{ij}.
\]
then for any $\gamma_j$ such that $(\tau_j)^{-1} \left\{ 32c_2 \gamma_j n^{-1} (\log(p) + \log(n)) \right\}^{1/2} \leq \gamma_j \leq c_3 / 6$, almost surely we have $\hat{\theta} = \theta$, when $n$ is sufficiently large.

Appendix B.: Technical proofs

B.1 Proof of Proposition 1

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

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

$$
\text{Cov}(Y_j, X_1 | X_2, \ldots, q) = 0, \quad j = 2, \ldots, p,
$$

(26)

and

$$
\text{Cov}(Y_1, X_1 | Y_A, X_2, \ldots, q) \neq 0, \quad \text{for any } A \subseteq \{2, \ldots, p\}.
$$

(27)

By the local Markov property (Spirtes et al., 2000), (27) implies that $X_1 \rightarrow Y_1$ in $\mathcal{E}(\bar{\theta})$. Suppose $Y_1$ is not a leaf node in $\mathcal{E}(\bar{\theta})$. Without loss of generality, assume that $Y_1$ is an unmediated parent of $Y_2$. Then $\text{Cov}(Y_2, X_1 | X_2, \ldots, q) = 0$ but $X_1 \rightarrow Y_1$ 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 $\mathcal{E}(\theta)$ then it must be a leaf node in $\mathcal{E}(\bar{\theta})$. In both $\mathcal{E}(\theta)$ and $\mathcal{E}(\bar{\theta})$, the parents and interventions of $Y_1$ can be identified by

$$
\mathbb{E}(Y_1 | Y_{\{2, \ldots, p\}}, X) = \mathbb{E}(Y_1 | Y_{PA(\theta) \setminus \{1\}}, X),
$$

and

$$
\mathbb{E}(Y_1 | Y_{\{2, \ldots, p\}}, X) = \mathbb{E}(Y_1 | Y_{PA(\bar{\theta}) \setminus \{1\}}, X) = \mathbb{E}(Y_1 | Y_{PA(\bar{\theta}) \setminus \{1\}}, X_{IN(\bar{\theta})}).
$$

Consequently, $Y_1$ has the same parents and interventions in $\mathcal{E}(\theta)$ and $\mathcal{E}(\bar{\theta})$.

The foregoing 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 $(Y_{\{2, \ldots, p\}}, X)$ and the sub-DAG of nodes $Y_2, \ldots, Y_p$. By induction, we remove the leafs in $\mathcal{E}(\theta)$ until it is empty, leading to $\mathcal{E}(\theta) = \mathcal{E}(\bar{\theta})$. Finally, $\theta = \bar{\theta}$ 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 2

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 $U$ is nilpotent in that $U^p = 0$. An application of the matrix series expansion
yields that \((I-U)^{-1} = I + U + ... + U^{p-1}\). Using the fact that \(V = W(I-U)^{-1}\) from (6), we have, for any \(1 \leq l, j \leq p\),

\[
V_{lj} = \sum_{k=1}^{p} W_{lk}(I_{lj} + U_{lj} + ... + (U^{p-1})_{lj})
\]

where \(U_{kj}\) is the \((k, j)th\) entry of \(U\). If \(V_{lj} \neq 0\), then there exists \(k\) such that \(W_{lk} \neq 0\) and \((U^{r})_{lj} \neq 0\) for some \(0 \leq r \leq p - 1\). If \(r = 0\), then we must have \(k = j\), and \(X_{l} \rightarrow Y_{j}\). If \(r > 0\), then \(X_{l} \rightarrow Y_{k}\) and \(Y_{k}\) is an ancestor of \(Y_{j}\).

**Proof of (B)**

First, for any leaf node variable \(Y_{k}\) by Assumption 1, there exists an instrument \(X_{l} \rightarrow Y_{j}\). If \(V_{lj} \neq 0\) 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 \(V_{lj} \neq 0\) and \(V_{lj'} = 0\) 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 \(U_{j'} \neq 0\) and \((U^{r})_{lj'} = 0\) for \(r > 1\). Then \(V_{lj} = W_{lj}U_{lj'} \neq 0\), a contradiction.

**B.3 Proof of Proposition 3**

Suppose \(Y_{k}\) is an unmediated parent of \(Y_{j}\). Let \(X_{j}\) be an instrument of \(Y_{k}\) in \(\mathcal{G}_{\ell'}\). Then there are two cases: (1) \(X_{j}\) intervenes on \(Y_{k}\) but does not intervene on \(Y_{j}\) namely \(X_{j} \rightarrow Y_{k}\) but \(X_{l} \rightarrow Y_{j}\); (2) \(X_{j}\) intervenes on \(Y_{k}\) and \(Y_{j}\) simultaneously, namely \(X_{j} \rightarrow Y_{k}\) and \(X_{l} \rightarrow Y_{j}\). For (1), \(V_{lj} = W_{lj}U_{lj'} \neq 0\). For (2), Assumption 1B implies that \(V_{lj} \neq 0\). This holds for every instrument of \(Y_{k}\) in \(\mathcal{G}_{\ell'}\), and the desired result follows.

Conversely, suppose for each instrument \(X_{j}\) of \(Y_{k}\) in \(\mathcal{G}_{\ell'}\), we have \(V_{lj} = 0\). Let \(X_{j'}\) be an instrument of \(Y_{k}\) in \(\mathcal{G}_{\ell'}\), which is also an instrument in \(\mathcal{G}_{\ell'}\). Then \(0 = V_{lj} = W_{lj}U_{lj'}\), which implies \(U_{lj} \neq 0\). This completes the proof.

**B.4 Proof of Theorem 4**

The proof proceeds in two steps: (A) we show that \(\{t: \widetilde{V}_{lj} \neq 0\} = \{t: V_{lj} \neq 0\}\) almost surely for \(1 \leq j \leq p\); and (B) we show that \(\mathcal{G}_{\ell} = \mathcal{G}_{\ell'}\) if \(\widetilde{V}\) satisfies the property in (A).

**Proof of (A)**

Let \(A_{j}^{l} = \{t: V_{lj} \neq 0\}\) and \(A_{jj}^{l} = \{t: \widetilde{V}_{lj} \geq \tau_{j}\}\) be the estimated support of penalized solution at the \(t\)-th iteration of Algorithm 1. For the penalized solution, define the false negative set \(FN_{j}^{l} = A_{j}^{l} \setminus A_{jj}^{l}\) and the false positive set \(FP_{j}^{l} = A_{jj}^{l} \setminus A_{j}^{l}\) for \(t \geq 0\). Consider a “good” event

\[
\mathcal{G}_{\ell} = \left\{\|X^T \hat{\mathcal{X}}_{\ell}/\alpha\|_{\infty} \leq 0.5\tau_{j}, \|\widetilde{V}_{lj} - V_{lj} / \alpha\|_{\infty} \leq 0.5\tau_{j}\right\}
\]
where $\hat{\xi}_j = Y_j - XV_j$, is the residual of the oracle least squares estimate $\hat{V}_j$, such that $A_j = \{j: \hat{V}_j \neq 0\}; 1 \leq j \leq p$. We shall show that $FN_j$ and $FP_j$ are eventually empty on event $\mathcal{E}$, which has a probability tending to one.

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

$$\left(\hat{V}_j - V_j \right)^T \left(-X^T (Y_j - X \hat{V}_j)/n + \gamma \tau \nabla \|\hat{V}_j \|_1, \nabla \|\hat{V}_j \|_1\right) \geq 0,$$

where $V_j$ is defined in (8). Plugging $\hat{\xi}_j = Y_j - X \hat{V}_j$ into the inequality and rearranging it, we have $\|X(\hat{V}_j - V_j)\|^2/n$ is no greater than

$$\left(\hat{V}_j - V_j \right)^T \left(X^T \hat{\xi}_j/n - \gamma \tau \nabla \|\hat{V}_j \|_1, \nabla \|\hat{V}_j \|_1\right) = \left(\hat{V}_{V_j} - V_{V_j} \right)^T \left(X^T \hat{\xi}_j/n - \gamma \tau \nabla \|\hat{V}_{V_j} \|_1, \nabla \|\hat{V}_{V_j} \|_1\right) + \left(\hat{V}_{V_j} - V_{V_j} \right)^T \nabla \|\hat{V}_{V_j} \|_1, \nabla \|\hat{V}_{V_j} \|_1\right),$$

where $X(\hat{V}_j - V_j)/n = 0$ has been used. Note that

$$\left(\hat{V}_{V_j} - V_{V_j} \right)^T \nabla \|\hat{V}_{V_j} \|_1, \nabla \|\hat{V}_{V_j} \|_1\right) = \left(\hat{V}_{V_j} - V_{V_j} \right)^T \nabla \|\hat{V}_{V_j} \|_1, \nabla \|\hat{V}_{V_j} \|_1\right).$$

Then (28) is no greater than

$$\|\hat{V}_{V_j} - V_{V_j} \|_1, \nabla \|\hat{V}_{V_j} \|_1\right) \left(\|X^T \hat{\xi}_j/n\|_{\infty} + \gamma \tau\right) + \left(\|X^T \hat{\xi}_j/n\|_{\infty} - \gamma \tau\right),$$

where $\Delta$ denotes the symmetric difference of two sets. Note that $\|X(\hat{V}_j - V_j)/n\|^2/n \geq 0$. Rearranging the inequality yields that

$$\left(\gamma \tau \|X^T \hat{\xi}_j/n\|_{\infty} + \|X^T \hat{\xi}_j/n\|_{\infty} - \gamma \tau\right) \|\hat{V}_{V_j} - V_{V_j} \|_1, \nabla \|\hat{V}_{V_j} \|_1\right).$$

On event $\mathcal{E}$, $\|X^T \hat{\xi}_j/n\|_{\infty} \leq \gamma \tau/2$, implying that

$$\|\hat{V}_{V_j} - V_{V_j} \|_1, \nabla \|\hat{V}_{V_j} \|_1\right) \leq 3 \|\hat{V}_{V_j} - V_{V_j} \|_1, \nabla \|\hat{V}_{V_j} \|_1\right).$$

Note that $|A_j \cup A_j^{-1}| \leq 2k_{max}$. By Assumption 2A and (29),
\begin{equation}
\begin{aligned}
c_i \| \tilde{V}_j^{(t)} - \tilde{V}_j^{\ell} \|_2^2 & \leq (\| X^T \tilde{\xi}_j / n \|_\infty + \gamma \tau) \| \tilde{V}_j^{(t)} - \tilde{V}_j^{A_j A_j^{-1}} \|_2^2 \\
& \quad + (\| X^T \tilde{\xi}_j / n \|_\infty - \gamma \tau) \| \tilde{V}_j^{A_j A_j^{-1}} - \tilde{V}_j^{A_j A_j^{-1}} \|_2^2 \\
& \leq (\| X^T \tilde{\xi}_j / n \|_\infty + \gamma \tau) \| \tilde{V}_j^{A_j A_j^{-1}} - \tilde{V}_j^{A_j A_j^{-1}} \|_2.
\end{aligned}
\end{equation}

By the Cauchy-Schwarz inequality,

\[
\| \tilde{V}_j^{A_j A_j^{-1}} - \tilde{V}_j^{A_j A_j^{-1}} \|_2 \leq \sqrt{A_j A_j^{-1}} \| \tilde{V}_j^{A_j A_j^{-1}} - \tilde{V}_j^{A_j A_j^{-1}} \|_2.
\]

Thus, \( c_i \| \tilde{V}_j^{(t)} - \tilde{V}_j^{\ell} \|_2 \leq 1.5 \gamma \tau \sqrt{2 \kappa_{\text{max}}} \) since \( |A_j A_j^{-1}| \leq |A_j| \leq 2 \kappa_{\text{max}} \) and \( \| X^T \tilde{\xi}_j / n \|_\infty \leq 0.5 \gamma \tau \). By the condition of Theorem 4, \( \| \tilde{V}_j^{(t)} - \tilde{V}_j^{\ell} \|_2 / \tau_j \leq \sqrt{\kappa_{\text{max}}} \), and \( \gamma \leq c_i / 6 \).

On the other hand, for any \( l \in \text{FP}^{(t)} = A_j A_j^{-1} \), we have \( \| \tilde{V}_j^{\ell} - \tilde{V}_j^{l} \|_2 = \| \tilde{V}_j^{(t)} \|_2 > \tau_j \). Thus, \( \sqrt{\text{FP}^{(t)}} \leq \| \tilde{V}_j^{(t)} - \tilde{V}_j^{\ell} \|_2 / \tau_j \leq \sqrt{\kappa_{\text{max}}} \), implying \( |A_j A_j^{-1}| = |A_j| + |\text{FP}^{(t)}| \leq 2 \kappa_{\text{max}} \) on \( \mathcal{E}_t \) for \( t \geq 1 \).

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

Now, suppose \( |\text{FN}^{(t)}| + |\text{FP}^{(t)}| \geq 1 \). For any \( l \in \text{FN}^{(t)} \cup \text{FP}^{(t)} \), by Assumption 3,

\[
\| \tilde{V}_j^{(t)} - \tilde{V}_j^{l} \|_2 \geq |\tilde{V}_j^{(t)}| - |\tilde{V}_j^{l}| - |\tilde{V}_j^{(t)} - \tilde{V}_j^{l}| \geq \tau_j - 0.5 \gamma \tau_j,
\]

so \( \sqrt{|\text{FN}^{(t)}| + |\text{FP}^{(t)}|} \leq \| \tilde{V}_j^{(t)} - \tilde{V}_j^{l} \|_2 / 0.5 \gamma \tau_j \). By (30) and the Cauchy-Schwarz inequality,

\[
c_i \| \tilde{V}_j^{(t)} - \tilde{V}_j^{\ell} \|_2 \leq 1.5 \gamma \tau \sqrt{|\text{FN}^{(t)}| + |\text{FP}^{(t)}|}.
\]

By conditions (1) and (2) for \( (\tau_j, \gamma_j) \) in Theorem 4, we have

\[
\sqrt{|\text{FN}^{(t)}| + |\text{FP}^{(t)}|} \leq \frac{\| \tilde{V}_j^{(t)} - \tilde{V}_j^{\ell} \|_2}{0.5 \gamma \tau_j} \leq \frac{3 \gamma \tau / c_i}{0.5 \sqrt{|\text{FN}^{(t)}| + |\text{FP}^{(t)}|}} \leq 0.5 \sqrt{|\text{FN}^{(t)}| + |\text{FP}^{(t)}|}.
\]

Hence, \( \sqrt{|\text{FN}^{(t)}| + |\text{FP}^{(t)}|} \leq (1/2)^t \sqrt{|\text{A}_j| + |A_j^{t-1}|} \). In particular, for \( t \geq 1 + [\log \kappa / \log 4] \),

\( |\text{FN}^{(t)}| + |\text{FP}^{(t)}| < 1 \) implying that \( \text{FN}^{(t)} = \emptyset \) and \( \text{FP}^{(t)} = \emptyset \).

Let \( t_{\text{max}} = 1 + [\log \kappa / \log 4] \). Then \( \text{FN}^{(t_{\text{max}})} = \text{FP}^{(t_{\text{max}})} = \emptyset \) on event \( \mathcal{E}_t \). By the condition of Theorem 4, we have \( \{ l : \tilde{V}_j^{(t_{\text{max}})} \neq 0 \} \subseteq \{ l : \tilde{V}_j^{\ell} \neq 0 \} \). To bound \( P( \cup_{l \neq 1} \mathcal{E}_t) \), let \( \eta = X^T (I - P_{A_j}) \xi_j \) and \( \eta' = n (X_j X_j^{-1})^T X_j^T \xi_j \), where \( \xi_j = (e_{v_1}, \ldots, e_{v_{\|l\|}})^T \). Then \( \eta \in \mathbb{R}^{d_j} \) and \( \eta' \in \mathbb{R}^{d_j} \) are Gaussian vectors with \( \text{Var}(\eta_j) \leq \Omega_j^1 |X_j^T X_j| \leq n \Omega_j^1 e_j \) and \( \text{Var}(\eta_j) \leq \Omega_j^1 |(X_j X_j^{-1})^{-1}| \leq n \Omega_j^{-1} e_j \).

Then
\[ 
P\left(\|X_\ell^n\|_\infty > 0.5r_j\right) = P\left(\|\eta\|_\infty > 0.5r_j\right) 
\leq \sum_{l=1}^q P(|\eta_l| \geq 0.5nr_j) 
\leq q \int_{1.25n^2r_j^2}^{\alpha} e^{-t^2/2} \sqrt{2\pi} dt 
\leq q \sqrt{\frac{\pi}{2}} \exp\left(-\frac{n\Omega_j^2 r_j^4}{8c_1}\right). 
\]

and similarly,
\[ 
P\left(\|V_\ell^n - V_\ell^n\|_\infty > 0.5r_j\right) \leq \sum_{l \in A_j} P(|\eta_l| > 0.5r_j) \leq \kappa \sqrt{\frac{\pi}{2}} \exp\left(-\frac{n\Omega_j^2 r_j^4}{8c_1}\right). 
\]

Under conditions for \((\tau_j, \gamma_j)\) in Theorem 4,
\[ 
P\left(\ell: V_{\ell,j}^{[\text{max}]} \neq 0\right) = \{\ell: V_{\ell,j} \neq 0\}; 1 \leq j \leq p \]
\[ 
\geq 1 - p \left(\frac{1}{p}\right)^{c_1} 
\geq 1 - p \left(\frac{1}{p}\right)^{c_1} 
\geq 1 - p \left(\frac{1}{p}\right)^{c_1} e^{-4(\log(q) + \log(n)) + \log(q) + e^{-144c_1^2c_2(\log(q) + \log(n)) + \log(k)}} \geq 1 - \frac{1}{\sqrt{\pi}} \rho^{-3} n^{-4}. 
\]

where \(c_1 < 6c_2\) is used in the last inequality. By the Borel-Cantelli lemma,
\[ 
\left\{\ell: V_{\ell,j}^{[\text{max}]} \neq 0\right\} = \{\ell: V_{\ell,j} \neq 0\}; 1 \leq j \leq p \text{ almost surely when } n \text{ is sufficiently large.}
\]

So far, we have \(\tilde{V}_j = V_{\ell,j}^{[\text{max}]} = \tilde{V}_j\); 1 \(\leq j \leq p\), almost surely. It remains to show that \(\tilde{V}_j\) 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),
\[ 
P\left(\tilde{V}_j \text{ is not a global minimizer of (7); } 1 \leq j \leq p\right) \leq 3 \exp\left(-2(\log(q) + \log(n))\right). 
\]

implying that almost surely \(\tilde{V}_j = V_{\ell,j}^{[\text{max}]} = \tilde{V}_j\) is a global minimizer of (7) when \(\rho\) is sufficiently large. This completes the proof.

**Proof of (B)**

Assume \(\tilde{V}\) satisfies the properties in (A). Then Propositions 2 and 3 holds true for \(\tilde{V}\). Clearly, we have \(\tilde{\mathcal{F}}_\alpha = \{(l, j): \tilde{V}_{\alpha} \neq 0 \text{ if } k = j \text{ or } (k, j) \in \tilde{\mathcal{E}}_\alpha\} = \mathcal{F}_\alpha \text{ whenever } \tilde{\mathcal{E}}_\alpha = \mathcal{E}_\alpha\). Thus, we only need to show \(\tilde{\mathcal{E}}_\alpha = \mathcal{E}_\alpha\). We shall prove this by induction.

Given \(\mathcal{G}^{\text{work}}\) and \(V^{\text{work}}\), the set of instruments on leaves is
\[ 
\mathcal{B} = \{l: l \text{ minimizes } \|V_{l,\alpha}^{\text{work}}\|_0 \text{ and } \|V_{l,\alpha}^{\text{work}}\|_0 > 0\} = \{\ell: \|V_{\ell,\alpha}^{\text{work}}\|_0 = 1\}.
\]
Hence, $X_j$ is an instrument of leaf variable $Y_k$ in $\mathcal{G}^{\text{work}}$, when $l \in \mathcal{S}$ and $k$ maximizes $|V^{\text{work}}_k|$. Hence, $\mathcal{S}$ is the index set of leaves in $\mathcal{G}^{\text{work}}$. By Proposition 3, all $(k, j)$ such that $Y_k$ is an unmediated parent of a peeled variable $Y_j$ are in $\mathcal{S}$, and can be identified by Assumption 1B. In model (1), after removing $Y_{\mathcal{S}}$, 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 $\mathcal{S}$, that contains all edges from an unmediated parent, so it suffices to recover $\mathcal{S}$. Consequently, $\mathcal{S}$, can be reconstructed by Step 9 and $\mathcal{S}$, can be recovered by Step 10. This completes the proof of (B).

**Lemma 11** Let

$$T_j = \frac{\mathbf{e}_j^\top (\mathbf{P}_{a_j} - \mathbf{P}_{b_j}) \mathbf{e}_j}{\sigma_j^2 |(\mathbf{I} - \mathbf{P}_{b_j})\mathbf{e}_j|/(n - |A_j|)}, \quad T_j^* = \frac{\mathbf{e}_j^\top (\mathbf{P}_{a_j} - \mathbf{P}_{b_j}) \mathbf{e}_j}{\sigma_j^2 |(\mathbf{I} - \mathbf{P}_{a_j})\mathbf{e}_j|/(n - |A_j|)},$$

where $A_j = \mathbf{P}_{a_j}(j) \cup \mathbf{IN}_j(j)$ and $B_j = (\mathbf{P}_{a_j}(j) \cup \mathbf{IN}_j(j))\mathbf{D}_j(j)$. Then $(T_1, \ldots, T_p)$ are independent and $(T_1^*, \ldots, T_p^*)$ are conditionally independent given $\mathbf{Z}$. Moreover,

$$T_j/(\,(|A_j| - |B_j|) \mid \mathbf{Z} - T), T_j/(\,(|A_j| - |B_j|) - F_{|A_j| - |B_j|, n - |A_j|}; 1 \leq j \leq p,$$

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

Proof of Lemma 11

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

$$\frac{\mathbf{e}_j^\top (\mathbf{P}_{a_j} - \mathbf{P}_{b_j}) \mathbf{e}_j}{\sigma_j^2 |(\mathbf{I} - \mathbf{P}_{b_j})\mathbf{e}_j|/(n - |A_j|)}, \quad \frac{\mathbf{e}_j^\top (\mathbf{I} - \mathbf{P}_{a_j}) \mathbf{e}_j}{\sigma_j^2 |(\mathbf{I} - \mathbf{P}_{a_j})\mathbf{e}_j|/(n - |A_j|)},$$

and they are independent, because $\mathbf{P}_{a_j} - \mathbf{P}_{b_j}$ and $\mathbf{I} - \mathbf{P}_{a_j}$ are orthonormal projection matrices, $\mathbf{e}_j \mid \mathbf{Z}_{a_j} - \mathcal{N}(\mathbf{0}, \sigma_j^2 \mathbf{I})$, 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 $t \mapsto \exp(it T_j/(\,(|A_j| - |B_j|)))$ is ( $t$ is the imaginary unit)

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

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, $T_j/(\,(|A_j| - |B_j|) - F_{|A_j| - |B_j|, n - |A_j|}; 1 \leq j \leq p).$ Similarly, we also have $T_j^*/(\,(|A_j| - |B_j|) - F_{|A_j| - |B_j|, n - |A_j|}; 1 \leq j \leq p)$. $F_{d_1, d_2}$.

Next, we prove independence for $\mathbf{T}$ and $\mathbf{T}^*$ via a peeling argument. Let $t = (t_1, \ldots, t_p)$. Let $Y_j$ be a leaf node of the graph $\mathcal{G}$. Then $\mathbf{T}_{-j} \mid \mathbf{Y}_{-j}, \mathbf{X}$ is deterministic, where $\mathbf{T}_{-j}$ is the subvector of $\mathbf{T}$ with the $j$th component removed. The characteristic function of $it^\top \mathbf{T}$ is
\[ \mathbb{E}\exp(t^\top T) = \mathbb{E}(\mathbb{E}(\exp(t^\top T_\cdot) \mid Y_\cdot, \mathbf{X})\exp(t^\top T_\cdot)) \]
\[ = \psi_{(d_j \to \emptyset) \cdots (d_1 \to \emptyset)}(t)\mathbb{E}\exp(t^\top T_\cdot), \]  

where \( \psi_{(d_j \to \emptyset) \cdots (d_1 \to \emptyset)} \) is the characteristic function of the F-distribution with degrees of freedom \( (|A| - |B|, n - |A|) \). Next, let \( Y_j \) be a leaf node of the graph \( G' \) and apply the law of iterated expectation again, where \( G' \) is the subgraph of \( G \) without node \( Y_j \). Repeat this procedure and after \( p \) steps \( \mathbb{E}\exp(t^\top T) = \prod_{j=1}^{p} \mathbb{E}\exp(t^\top T_j) \), which implies \( (T_1, \ldots, T_p) \) are independent. Similarly, \( T^* \) also has independent components given \( \mathbf{Z} \) and has the same distribution as \( T \). This completes the proof.

### B.5 Proof of Theorem 5

Let \( L_r(\delta, \Sigma) \) denote the likelihood ratio given \( \delta \) and \( \Sigma \). Then \( L_r = L_r(\widehat{\delta}, \widehat{\Sigma}) \).

**Proof of (A)**

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

\[
\frac{\mathbb{P}(L_r \leq x) - \mathbb{P}(L_r(\delta, \Sigma) \leq x)}{\mathbb{E}[I(L_r \leq x) - I(L_r(\delta, \Sigma) \leq x)]} \\
 \leq \mathbb{E}[I(L_r \leq x) - I(L_r(\delta, \Sigma) \leq x)] \\
 = \mathbb{E}[I(L_r \leq x, \widehat{\delta} \neq \delta) + I(L_r(\delta, \Sigma) \leq x)(I(\widehat{\delta} = \delta) - 1)] \\
 \leq 2\mathbb{P}(\widehat{\delta} \neq \delta). \tag{33}
\]

From (14), \( L_r^*(\delta, \Sigma^*) = \sum_{j \notin \{i, j \neq \delta \}} T_j^* \). By Lemma 11,

\[ \mathbb{P}(L_r(\delta, \Sigma) \leq x) = \mathbb{P}(L_r^*(\delta, \Sigma^*) \leq x \mid \mathbf{Z}). \]

Note that \( L_r^* = L_r^*(\widehat{\delta}^*, \Sigma^*) \). Then for any real number \( x \),

\[
\frac{\mathbb{P}(L_r^* \leq x \mid \mathbf{Z}) - \mathbb{P}(L_r^*(\delta, \Sigma^*) \leq x \mid \mathbf{Z})}{\mathbb{P}(L_r^*(\delta, \Sigma^*) \leq x \mid \mathbf{Z})} \\
 = \frac{\mathbb{P}(L_r^* \leq x, \widehat{\delta}^* \neq \delta \mid \mathbf{Z}) + \mathbb{P}(L_r^* \leq x, \widehat{\delta}^* = \delta \mid \mathbf{Z}) - \mathbb{P}(L_r^*(\delta, \Sigma^*) \leq x \mid \mathbf{Z})}{\mathbb{P}(L_r^*(\delta, \Sigma^*) \leq x \mid \mathbf{Z})} \\
 \leq \frac{\mathbb{P}(L_r^* \leq x, \widehat{\delta}^* \neq \delta \mid \mathbf{Z})}{\mathbb{P}(L_r^*(\delta, \Sigma^*) \leq x \mid \mathbf{Z})} + \frac{\mathbb{P}(L_r^*(\delta, \Sigma^*) \leq x \mid \mathbf{Z})}{\mathbb{P}(L_r^*(\delta, \Sigma^*) \leq x \mid \mathbf{Z})} \\
 \leq 2\mathbb{P}(\widehat{\delta}^* \neq \delta \mid \mathbf{Z}). \tag{34}
\]

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

\[ \sup_{x \in \mathbb{R}} |\mathbb{P}(L_r \leq x) - \mathbb{P}(L_r^* \leq x \mid \mathbf{Z})| \leq 2\mathbb{P}(\widehat{\delta}^* \neq \delta) + 2\mathbb{P}(\widehat{\delta}^* \neq \delta \mid \mathbf{Z}). \]

By Theorem 4, we have \( \mathbb{P}(\widehat{\delta}^* \neq \delta) \to 0 \), which holds uniformly for all \( \mathbf{\theta} \) which satisfy the Assumptions 1-4. For \( \mathbb{P}(\widehat{\delta}^* \neq \delta \mid \mathbf{Z}) \), the error terms of (7) in the perturbed data \( \mathbf{Z}^* \) are rescaled with \( \Omega^*_j + \delta_j^* \leq 2\Omega^*_j \). By Theorem 4, \( \mathbb{P}(\widehat{\delta}^* \neq \delta) = \mathbb{E}\mathbb{P}(\widehat{\delta}^* \neq \delta \mid \mathbf{Z}) \to 0 \), which implies \( \mathbb{P}(\widehat{\delta}^* \neq \delta \mid \mathbf{Z}) \to 0 \) as \( n \to \infty \) by the Markov inequality. Consequently,
Let \( P \) be a sequence of \( \ell = \sum \alpha_t \delta_t \) such that \( P \) satisfies Hypothesis \( H_0 \). By Definition 7, \( P \) is distributional with \( \mathcal{P} \). Hence, 

\[ \lim \sup_{n \to \infty} P_{\alpha} \left(\forall \theta \in \mathcal{P} \right) \geq \lim \sup_{n \to \infty} P_{\alpha} \left(\forall \theta \in \mathcal{P} \right) < \infty \]

This completes the proof.

**B.6 Proof of Proposition 6**

Since \( \mathbb{P}(\mathcal{C} = \mathcal{S}) \to 1 \), it suffices to consider \( L_\theta(\mathcal{C}, \mathcal{S}) \). For \( \lvert \theta \rvert = 0 \), we have \( \mathbb{P}(L_\theta = 0) \to 1 \).

Now, assume \( \lvert \theta \rvert > 0 \). Then

\[ 2L_\theta(\mathcal{C}, \mathcal{S}) = \sum_{\{j: D_{\theta}(j) = 0\}} \frac{e_j(P_{\alpha} - P_{\beta})e_j}{\sigma_j^2} + \sum_{\{j: D_{\theta}(j) = 0\}} \left(\frac{\sigma_j^2}{\sigma_j^2} - 1\right) e_j(P_{\alpha} - P_{\beta})e_j. \]

To derive the asymptotic distribution of \( R_1 \), we apply the strategy with law of iterated expectation as in the proof of Lemma 11. Then we have \( \{e_j(P_{\alpha} - P_{\beta})e_j/\sigma_j^2\}_{\theta \neq 0} \) are independent. Therefore, \( R_1 \sim \chi^2_{(\eta - 1)} \). To bound \( R_2 \), we apply Lemma 1 of Laurent and Massart (2000). By Assumption 5,

\[ \mathbb{P}\left( \max_{\{j: D_{\theta}(j) \neq 0\}} \left| \frac{\sigma_j^2}{\sigma_j^2} - 1 \right| \geq 4 \frac{\log(|\theta|)}{(1 - \rho)n} + 8 \frac{\log(|\theta|)}{(1 - \rho)n} \right) \leq 2 \exp\left( - \log(|\theta|) \right). \]

Hence,

\[ \max_{\{j: D_{\theta}(j) \neq 0\}} \left| \frac{\sigma_j^2}{\sigma_j^2} - 1 \right| \leq 8 \sqrt{\log(|\theta|)}/(1 - \rho)n \]with probability tending one. Thus

\[ |R_2| \leq \mathcal{O}\left( \left| \mathcal{C} \right| \sqrt{\frac{\log(|\theta|)}{n}} \right). \]
Consequently, the desired result follows.

### B.7 Proof of Proposition 7

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

$$
L(\theta^{(n)}, \Sigma) - L(\theta^*, \Sigma) = \sum_{\{j \in D_0(j) \neq \emptyset\}} \left( \sqrt{n}\eta_j \Delta_j - \frac{1}{2} \sqrt{n} \Delta_j \left(n^{-1} Y_{D_0(j)}^T Y_{D_0(j)} \right)^{1/2} \right).
$$

where $\eta_j = (n^{-1/2} Y_{D_0(j)}^T e_j, 0_{D_0(j)})$ is a $p$-vector. It suffices to consider $\Pr(\delta, \Sigma)$, since $\Pr(\delta) \to 1$. Under null hypothesis $H_0$, $2Lr = \sum_{j \in D_0(j) \neq \emptyset} T_j$, where by Proposition 6 we have $T_j = \sum_{r=1}^{\alpha_{D_0(j)}(j)} (q_{j,r} e_j)^2 + \alpha_0(1)$ with $p_{j,r} = \sum_{\alpha_{D_0(j)}(j)} q_{j,r} k_{j,r}$. Letting $Q_j = (q_{j,1}, \ldots, q_{j,\alpha_{D_0(j)}})$ be an $n \times 1D\alpha(j)$ matrix, then

$$
\begin{bmatrix} Q_j^T e_j & \eta_j \\ Y_{D_0(j)}^T \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} I_{p_{D_0(j)}} & n^{-1/2} Q_j^T Y_{D_0(j)} \\ n^{-1/2} Y_{D_0(j)}^T Q_j & n^{-1} Y_{D_0(j)}^T Y_{D_0(j)} \end{bmatrix}, 2Lr \right).
$$

where $Z_j \sim \mathcal{N}(0, I_{p_{D_0(j)}})$. By Lemma 3 of Li et al. (2020), $\|Q_j Y_{D_0(j)} \Delta_j\|^2 \geq c_j \|\Delta_j\|^2$ with probability tending to one. Therefore, a peeling argument yields

$$
\beta(\theta^*, \Delta) \geq \Pr(\|Z + c\Delta\|^2 \geq \chi^2_{p_{D_0(j)}}).
$$

where $Z \sim \mathcal{N}(0, I_{\alpha_{D_0(j)}})$. Similarly, as $|\|D\alpha|| \to \infty$,

$$
\beta(\theta^*, H) \geq \Pr(\|Z + c\Delta\|^2 \geq \chi^2_{p_{D_0(j)}} - |\|D\alpha|||) \to \Pr(\|Z + c\Delta\|^2 \geq \chi^2_{|\|D\alpha|||}) = \Pr(Z > z_{1-\alpha} - c \|\Delta\| / \sqrt{2} |\|D\alpha|||).
$$

This completes the proof.
B.8 Proof of Proposition 8

By Proposition 7, for fixed \(||\mathcal{A}|| = |\mathcal{R}| > 0\), \(\rho(\theta', \Delta)\) equals to
\[
P(\text{Pval}_i < a, \ldots, \text{Pval}_p < a) \geq 1 - \sum_{k = 1}^{p} P(\text{Pval}_k \leq a) \geq 1 - |\mathcal{R}| P(Z^2_i \leq x^2_{1,\ldots,n}).
\]

where \(Z_i \sim N(\delta^l, \max_1 \Omega_i, 1)\). Then \(P(Z^2 \leq x) \leq P(Z \leq -\mu + \sqrt{x}) \leq \frac{1}{\sqrt{2\pi}} e^{-(\mu - \sqrt{x})^2/2} \) for \(Z \sim N(\mu, 1)\) and any \(0 \leq x < \mu^2 \) and \(\mu > 0\) (Small, 2010). Hence, as \(||\mathcal{A}|| = |\mathcal{R}| \rightarrow \infty\),
\[
\rho(\theta', \Delta) \geq 1 - |\mathcal{R}| P(Z^2_i \leq x^2_{1,\ldots,n}) \geq 1 - \sum_{k = 1}^{p} P(\text{Pval}_k \leq a) \geq 1 - |\mathcal{R}| P(Z^2_i \leq x^2_{1,\ldots,n}).
\]

This completes the proof.

B.9 Proof of Theorem 9

The proof proceeds in two steps to show that

A. \(\{l: |\mathcal{N}_l| \geq \tau'\} \subseteq \{l: \bar{V}_l \neq 0\} \subseteq \{l: \mathcal{N}_l \neq 0\}\) for \(1 \leq j \leq p\) almost surely, and

B. \(\mathcal{G}_* = \mathcal{G}_+\) when \(\bar{V}\) satisfies the property in (A).

**Proof of (A)**

This proof is similar to that of Theorem 4 part (A). To proceed, let \(\mathcal{A}' = \{l: V_l \neq 0\}\), \(A'_i = \{l: |\mathcal{N}_l| > \tau'_i\}\), and \(A_i^{(0)} = \{l: ||\bar{V}|| > \tau'_i\}\). Then we define the false negative set \(\text{FN}^{(0)} = A_i^{(0)} \setminus A_i^{(0)}\), and the false positive set \(\text{FP}^{(0)} = A_i^{(0)} \setminus A_i^{(0)}\) for \(t \geq 0\). Consider
\[
\mathcal{G}_j = \{|\bar{V}_j - X\bar{V}_j|_\infty \leq 0.5\tau'_j\} \cap \{|\bar{V}_j|_\infty \leq 0.5\tau'_j\},
\]

where \(\bar{V}_j = Y_j - X\bar{V}_j\), for \(1 \leq j \leq p\). Again, we shall show that \(\text{FN}^{(0)}\), and \(\text{FP}^{(0)}\), are eventually empty on event \(\mathcal{G}_j\), which has a probability tending to one.

Assume \(|\mathcal{A}'_i \cup A_i^{(0)}| \leq 2\kappa_{\max}\). Then (28), (29), and (30) remain true on \(\mathcal{G}_j\). As a result, \(|\mathcal{A}'_i \cup A_i^{(0)}| \leq 2\kappa_{\max}\) on \(\mathcal{G}_j\), for \(t \geq 1\).

To estimate the number of iterations required for termination, it suffices to prove that \(|\text{FN}^{(0)}| + |\text{FP}^{(0)}| < 1\) eventually. Suppose \(|\text{FN}^{(0)}| + |\text{FP}^{(0)}| \geq 1\). For any \(l \in \text{FN}^{(0)} \cup \text{FP}^{(0)}\), by Assumption 5,
\[
|\bar{V}_i^j - \bar{V}_j| \geq |\bar{V}_i^j - V_i^j| - |\bar{V}_j| \geq \min(\tau'_j - \tau, \tau). \]


we have

\[ \sqrt{\mathbb{E}|N| + \mathbb{E}|P|} \leq \sqrt{\mathbb{E}|X| \mathbb{E}|Y|} \leq \min(\tau_0, \tau_1). \]

Letting \( \kappa_j = |A_i| \), by (30) and the Cauchy-Schwarz inequality,

\[ \epsilon \|\tilde{V}\| - \tilde{V}_j \| \leq 1.5\tau_0 \|A_i \| A_i \|^{-1} \| + 1.5\tau_0 \|A_i \|^{-1} \| + \|A_i \|^{-1} \| + (\kappa_j - \kappa_j'). \]

By the conditions for \( (\tau_k, \gamma_j) \) in Theorem 9, we have

\[ \sqrt{\mathbb{E}|N| + \mathbb{E}|P|} \leq \frac{\|\tilde{V}\| - \tilde{V}_j \|}{\min(\tau_0, \tau_1)} \leq \frac{1.5\tau_0 \|A_i \| A_i \|^{-1} \| + 1.5\tau_0 \|A_i \|^{-1} \| + \|A_i \|^{-1} \| + (\kappa_j - \kappa_j')}{\epsilon \min(\tau_0, \tau_1, \tau_2)} \leq 0.5\sqrt{\mathbb{E}|N| + \mathbb{E}|P|} + 0.25. \]

Hence, \( \sqrt{\mathbb{E}|N| + \mathbb{E}|P|} \leq (1/2)^t \|A_i \| + |A_i| \| + 0.5 \). For \( t \geq 1 + \lceil \log \epsilon /\log 4 \rceil \), we have \( \mathbb{E}|N| + \mathbb{E}|P| < 1 \) implying that \( \mathbb{E}|N| = \emptyset \) and \( \mathbb{E}|P| = \emptyset \).

Under the conditions for \( (\tau_k, \gamma_j) \) in Theorem 9, using the same argument in Proof of Theorem 4 we have

\[ P\{l: |V_j| > \tau_j \} \leq \{l: |\tilde{V}_j| > \tau_j \} \leq \{l: |\tilde{V}_j| = 0\} \leq 1 \leq |j| \leq \rho \geq 1 - \frac{\rho \gamma}{\mathbb{V}^2 \rho - 3^n - 4}. \]

By Borel-Cantelli lemma, almost surely we have

\[ \{l: |V_j| > \tau_j \} \subseteq \{l: |\tilde{V}_j| > \tau_j \} \subseteq \{l: |\tilde{V}_j| = 0\} \quad j = 1, \ldots, \rho. \]

when \( \rho \) is sufficiently large. Thus, we have \( \{l: |\tilde{V}_j| = 0\} = \{l: |\tilde{V}_j| > \tau_j \} \) whenever \( \kappa_j = \|f: |\tilde{V}_j| > \tau_j \| \); \( j = 1, \ldots, \rho \). The desired result follows.

**Proof of (B)**

This is the same as the Proof of Theorem 4 part (B).

### B.10 Proof of Theorem 10

By Theorem 4, it suffices to consider the event \( \{\tilde{z}_i = \tilde{y}_i\} \). Then with \( X \) being replaced by \( (Y_{\text{agg}, \text{agg}}, X_{\text{agg}, \text{agg}}) \), this proof is almost the same as that of Theorem 4 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

\[
\gamma_j' \in \left\{ \log(\max_{l,j} X_l Y_j), \ldots, 0.05 \log(\max_{l,j} X_l Y_j) \right\} \quad (100 \text{ equally spaced values}).
\]

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 \( \lambda \) sequence of 2SPLS (Chen et al., 2018) is set in the same way as the \( \gamma \) 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 \( U \), 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.
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

Angrist JD, Imbens GW, and Rubin DB. Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434):444–455, 1996.

Bergmann TO and Hartwigsen G Inferring causality from noninvasive brain stimulation in cognitive neuroscience. Journal of Cognitive Neuroscience, 33(2):195–225, 2021. [PubMed: 32530381]

Bickel PJ, Ritov Y, and Tsybakov AB. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 37(4):1705–1732, 2009.

Brecheny P and Huang J Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. The Annals of Applied Statistics, 5(1):232–253, 2011. [PubMed: 22081779]

Breiman L 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.

Brown BC and Knowles DA Phenome-scale causal network discovery with bidirectional mediated Mendelian randomization. bioRxiv, 2020.

Bu G Apolipoprotein E and its receptors in Alzheimer’s disease: pathways, pathogenesis and therapy. Nature Reviews Neuroscience, 10(5):333–344, 2009. [PubMed: 19339974]

Candes E, Fan Y, Janson L, and Lv J 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 C, Ren M, Zhang M, and Zhang D. A two-stage penalized least squares method for constructing large systems of structural equations. Journal of Machine Learning Research, 19(1):40–73, 2018.

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

Efron B, Hastie T, Johnstone I, and Tibshirani R Least angle regression. The Annals of Statistics, 32(2):407–499, 2004.
Fan J, Xue L, and Zou H. Strong oracle optimality of folded concave penalized estimation. The Annals of Statistics, 42(3):819–849, 2014. [PubMed: 25598560]

Friedman JH, Hastie T, and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22, 2010. [PubMed: 20808728]

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

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

Grosse-Wentrup M, Janzing D, Siegel M, and Schölkopf B. Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. NeuroImage, 125:825–833, 2016. [PubMed: 26518633]

Heinze-Deml C, Maathuis MH, and Meinshausen N. Causal structure learning. Annual Review of Statistics and Its Application, 5:371–391, 2018.

Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, and Linsley PS. Expression profiling reveals off-target gene regulation by RNAi. Nature Biotechnology, 21(6):635–637, 2003.

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

Julia T and Goate AM. Genetics of β-amyloid precursor protein in Alzheimer’s disease. Cold Spring Harbor Perspectives in Medicine, 7(6), 2017.

Kanehisa M and Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 28(1):27–30, 2000. [PubMed: 10592173]

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

Kleiner A, Talwalkar A, Sarkar P, and Jordan MI. The big data bootstrap. In Proceedings of the 29th International Conference on International Conference on Machine Learning, pages 1787–1794, 2012.

Kremer A, Louis JV, Jaworski T, and Van Leuven F. GSK3 and Alzheimer’s disease: facts and fiction. Frontiers in Molecular Neuroscience, 4:17, 2011. [PubMed: 21904524]

Kulkarni MM, Booker M, Silver SJ, Friedman A, Hong P, Perrimon N, and Mathey-Prevot B. Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays. Nature Methods, 3(10):833–838, 2006. [PubMed: 16964256]

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

Li C, Shen X, and Pan W. Likelihood ratio tests for a large directed acyclic graph. Journal of the American Statistical Association, 115(531):1304–1319, 2020. [PubMed: 32973370]

Liu Z, Zhang M, Xu G, Huo C, Tan Q, Li Z, and Yuan Q. Effective connectivity analysis of the brain network in drivers during actual driving using near-infrared spectroscopy. Frontiers in Behavioral Neuroscience, 11:211, 2017. [PubMed: 29163083]

Lockhart R, Taylor J, Tibshirani RJ, and Tibshirani R. A significance test for the lasso. The Annals of Statistics, 42(2):413–468, 2014. [PubMed: 25574062]

Loh P-L and Wainwright MJ. Support recovery without incoherence: A case for nonconvex regularization. The Annals of Statistics, 45(6):2455–2482, 2017.

Luo R and Zhao H. Bayesian hierarchical modeling for signaling pathway inference from single cell intervention data. The Annals of Applied Statistics, 5(2A):725–745, 2011. [PubMed: 22162986]

Matsui T, Ingelsson M, Fukumoto H, Ramasamy K, Kowa H, Froesch MP, Irizarry MC, and Hyman BT. Expression of APP pathway mRNAs and proteins in Alzheimer’s disease. Brain Research, 1161:116–123, 2007. [PubMed: 17586478]

Meinshausen N and Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):417–473, 2010.
Molstad AJ, Sun W, and Hsu L. 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. [PubMed: 34413922]

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

Oates CJ, Smith JQ, and Mukherjee S. Estimating causal structure using conditional DAG models. Journal of Machine Learning Research, 17(1):1880–1903, 2016.

Pamfil R, Sriwattanaworachai N, Desai S, Pilgerstorfer P, Georgatzis K, Beaumont P, and Aragam B. DYNOTEARS: Structure learning from time-series data. In International Conference on Artificial Intelligence and Statistics, pages 1595–1605. PMLR, 2020.

Pearl J. Causality. Cambridge University Press, 2009.

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

Peters J, Mooij JM, Janzing D, and Schölkopf B. Causal discovery with continuous additive noise models. Journal of Machine Learning Research, 15(1):2009–2053, 2014.

Peters J, Bühlmann P, and Meinshausen N. 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.

Rajendran G, Kivva B, Gao M, and Aragam B. 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.

Rolland P, Cevher V, Kleindessner M, Russell C, Janzing D, Schölkopf B, and Locatello F. Score matching enables causal discovery of nonlinear additive noise models. In International Conference on Machine Learning, pages 18741–18753. PMLR, 2022.

Rothenhoäusler D, Bühlmann P, and Meinshausen N. Causal Dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. The Annals of Statistics, 47(3):1688–1722, 2019.

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

Sachs K, Perez O, Pe’er D, Lauffenburger DA, and Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721):523–529, 2005. [PubMed: 15845847]

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

Shen X, Pan W, and Zhu Y. Likelihood-based selection and sharp parameter estimation. Journal of the American Statistical Association, 107(497):223–232, 2012. [PubMed: 22736876]

Shen X, Pan W, Zhu Y, and Zhou H. On constrained and regularized high-dimensional regression. Annals of the Institute of Statistical Mathematics, 65(5):807–832, 2013. [PubMed: 24465052]

Shi C, Song R, Chen Z, and Li R. Linear hypothesis testing for high dimensional generalized linear models. The Annals of Statistics, 47(5):2671–2703, 2019. [PubMed: 31534282]

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

Shojaie A and Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika, 97(3):519–538, 2010. [PubMed: 22434937]

Small CG. Expansions and Asymptotics for Statistics. Chapman and Hall/CRC, 2010.

Spirtes P, Glymour C, and Scheines R. Causation, Prediction, and Search. MIT Press, 2000.

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

Su JH, Zhao M, Anderson AJ, Srinivasan A, and Cotman CW. Activated caspase-3 expression in Alzheimer’s and aged control brain: correlation with Alzheimer pathology. Brain Research, 898(2):350–357, 2001. [PubMed: 11306022]
Teumer A. Common methods for performing Mendelian randomization. Frontiers in Cardiovascular Medicine, 5:51, 2018. [PubMed: 29892602]

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

Van de Geer S and Bühlmann P. \( \ell_0 \)-penalized maximum likelihood for sparse directed acyclic graphs. The Annals of Statistics, 41(2):536–567, 2013.

Viinikka J, Hyttinen A, Pensar J, and Koivisto M. Towards scalable Bayesian learning of causal DAGs. In Proceedings of the 34th International Conference on Neural Information Processing Systems, pages 6584–6594, 2020.

Wasserman L and Roeder K. High-dimensional variable selection. The Annals of Statistics, 37(5A):2178–2201, 2009. [PubMed: 19784398]

Yuan Y, Shen X, Pan W, and Wang Z. Constrained likelihood for reconstructing a directed acyclic Gaussian graph. Biometrika, 106(1):109–125, 2019. [PubMed: 30799877]

Zhang Y, Wainwright MJ, and Jordan MI. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In Conference on Learning Theory, pages 921–948. PMLR, 2014.

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

Zheng X, Aragam B, Ravikumar P, and Xing EP. 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.

Zhu Y, Shen X, and Pan W. On high-dimensional constrained maximum likelihood inference. Journal of the American Statistical Association, 115(529):217–230, 2020. [PubMed: 32788818]
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 $X$ are not displayed.
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.
Figure 5:
Rejection rates for $H_0: U_{1,20} = 0$ versus $H_a: U_{1,20} = 1/\sqrt{n}$ by DP inference and structure learning. For structure learning, $H_0$ is rejected if $\bar{U}_{1,20} \neq 0$. 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.
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).
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 $\alpha = 0.05$. 

(a) AD-MCI

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