Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Discovery and inference of a causal network with hidden confounding*

Li Chen, Chunlin Li, Xiaotong Shen & Wei Pan

To cite this article: Li Chen, Chunlin Li, Xiaotong Shen & Wei Pan (04 Oct 2023): Discovery and inference of a causal network with hidden confounding*, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2261658

To link to this article: https://doi.org/10.1080/01621459.2023.2261658

Discovery and inference of a causal network with hidden confounding*

Li Chen[a] Chunlin Li[b,#] Xiaotong Shen[a] Wei Pan[c]

[a]School of Statistics, University of Minnesota, Minneapolis, MN 55455.
[b]Department of Statistics, Iowa State University, Ames, IA 50011.
[c]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455.

*L. Chen and C. Li contributed equally. Research is supported in part by NSF grant DMS-1952539, NIH grants R01GM113250, R01GM126002, R01AG065636, R01AG074858, R01AG069895, U01AG073079. The authors report there are no competing interests to declare. The authors thank the editor, the associate editor, and three anonymous referees for their helpful comments and suggestions. C. Li would like to thank R. Oliver VandenBerg for the suggestions on writing.
#Email: chunlin@iastate.edu.

Abstract

This article proposes a novel causal discovery and inference method called GrIVET for a Gaussian directed acyclic graph with unmeasured confounders. GrIVET consists of an order-based causal discovery method and a likelihood-based inferential procedure. For causal discovery, we generalize the existing peeling algorithm to estimate the ancestral relations and candidate instruments in the presence of hidden confounders. Based on this, we propose a new procedure for instrumental variable estimation of each direct effect by separating it from any mediation effects. For inference, we develop a new likelihood ratio test of multiple causal effects that is able to account for the unmeasured confounders. Theoretically, we prove that the proposed method has desirable guarantees, including robustness to invalid instruments and uncertain interventions, estimation consistency, low-order polynomial time complexity, and validity of asymptotic inference. Numerically, GrIVET performs well and compares favorably against state-of-the-art competitors.

Furthermore, we demonstrate the utility and effectiveness of the proposed method through an application inferring regulatory pathways from Alzheimer's disease gene expression data.

Keywords: Causal discovery, Gaussian directed acyclic graph, Invalid instrumental variables, Uncertain interventions, Simultaneous inference, Gene regulatory network.

1 Introduction

Understanding causal relations is part of the foundation of intelligence. A directed acyclic graph (DAG) is often used to describe the causal relations among multiple interacting units (Pearl, 2009). Unlike classical causal inference tasks where the DAG is determined a priori, causal discovery aims to learn a graphical representation from data. It is useful for forming data-driven conjectures about the underlying mechanism of a complex system, including gene networks (Sachs et al., 2005), functional brain networks (Liu et al., 2017), manufacturing pipelines (Kertel et al., 2022), and dynamical systems (Li et al., 2020b). In such a situation, randomized experiments are usually unethical or infeasible, and unmeasured confounders commonly arise in practice. The presence of latent confounders can bias the causal effect estimation and even distort causal directions, making causal discovery challenging. To treat latent confounders, we use additive interventions as instrumental variables (IVs), which are well-developed in conventional causal inference (Angrist et al., 1996) yet are less explored in causal discovery of a large-scale network. In this article, we focus on a Gaussian DAG model with hidden confounders and develop methods that integrate the discovery and inference of causal relations within the framework of uncertain additive interventions (the targets of interventions are unknown).

Causal discovery has been extensively studied (Zheng et al., 2018; Aragam et al., 2019; Gu et al., 2019; Lee and Li, 2022; Zhao et al., 2022; Li et al., 2023b); see Drton and Maathuis (2017); Heinze-Deml et al. (2018); Glymour et al. (2019); Vowels et al. (2021) for comprehensive reviews. For observational data (without external interventions), some methods are able to treat hidden confounding by either (a) producing less informative discoveries, like a partial ancestral graph (Colombo et al., 2012) rather than a DAG, or (b) employing a certain deconfounding

strategy (Frot et al., 2019; Shah et al., 2020) based on the pervasive confounding assumption. However, the former may not reveal essential information, such as causal directions, while the latter can be inconsistent in low-dimensional situations and may not necessarily outperform the naive regression (Grimmer et al., 2020). Thus, external interventions are useful to provide more information about causal relations while relaxing the requirements on latent confounding.

As an example of external (additive) interventions, IVs have been well developed in conventional causal inference to tackle unmeasured confounding; see Lousdal (2018) for a survey. In a classical bivariate setting where the causal direction is known, an IV is required to influence the response variable only through the cause variable, which is often fragile in practice (Murray, 2006). For instance, genetic variants like single nucleotide polymorphisms (SNPs) are used as IVs in Mendelian randomization (MR) analysis to discover putative causal genes of complex traits, where the IV conditions are commonly violated due to the (horizontal) pleiotropy. Remedying these invalid IVs has been the subject of recent work in causal inference (Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2019; Burgess et al., 2020). The discussion of IV estimation in graphical modeling, however, remains limited. The methods of Oates et al. (2016); Chen et al. (2018) estimate the graph given valid IVs, while the work of Li et al. (2023a) propose the peeling algorithm to construct the DAG in the case of uncertain interventions and invalid IVs. None of these methods permit latent confounding. A recent work (Xue and Pan, 2020) discusses causal discovery of a bivariate mixed effect graph where confounders and invalid IVs are allowed, but it remains unclear how to extend it to a large-scale causal network.

Moreover, despite the progress in causal discovery, inference about the discovered relations is often regarded as a separate task and has received less attention in the literature. Notable exceptions include recent advances in graphical modeling (Janková and van de Geer, 2018; Li et al., 2020a; Shi et al., 2023; Wang et al., 2023) and mediation analysis (Chakrabortty et al., 2018; Shi and Li, 2021; Li et al., 2022); however, these methods cannot account for latent confounders. Indeed, due to unmeasured confounding, the probability distribution of observed variables is no longer locally Markovian with respect to the DAG (Pearl, 2009),

rendering these approaches inappropriate. Consequently, there is a pressing need for new inference methodologies.

This article contributes to the following aspects.

- For modeling, we establish the identifiability conditions for a Gaussian DAG with latent confounders utilizing additive interventions. To our knowledge, this result is the first of its kind. Importantly, the conditions allow the interventions to have unknown and multiple targets, which is suitable for multivariate causal analysis (Murray, 2006).
- For methodology, we develop a novel method named the Graphical Instrumental Variable Estimation and Testing (GrIVET), integrating order-based causal discovery and likelihood-based inference. For causal discovery, we estimate the ancestral relations and candidate IVs with a modified peeling algorithm to treat unmeasured confounding. On this basis, we propose a sequential procedure to estimate each direct effect using IVs, where a working response regression is used to separate the direct effect from the mediation effects. Regarding inference, we develop a new likelihood ratio test of multiple causal effects to account for unmeasured confounders.
- For theory, we show that GrIVET enjoys desired guarantees. In particular, it consistently estimates the DAG structure and causal effects even when some interventions do not meet the IV criteria. As for computation, only $O((p+|\mathcal{E}^+|)\times\log(s)\times(q^3+nq^2))$ operations are required almost surely, where p and q are the numbers of primary and intervention variables, s is sparsity, $|\mathcal{E}^+|$ is the size of the ancestral relation set, and n is the sample size. Moreover, under the null hypothesis, we establish the convergence of the likelihood ratio statistic to the null distribution in high-dimensional situations, ensuring the validity of asymptotic inference.
- The simulation studies and an application to the Alzheimer's Disease Neuroimaging Initiative dataset demonstrate the utility and effectiveness of the proposed methods. The implementation of GrIVET is available at https://github.com/chunlinli/grivet.

The rest of the article is structured as follows. Section 2 introduces a linear structural equation model with hidden confounders and establishes its identifiability. Section 3 presents a novel order-based method for causal discovery and effect estimation. Section 4 develops a likelihood ratio test for simultaneous inference of causal effects. Section 5 provides theoretical justification of the proposed method. Section 6 performs simulation studies, followed by an application to infer gene pathways with gene expression and SNP data. Finally, Section 7 concludes the article. The Appendix contains supporting lemmas, while the Supplementary Materials include illustrative examples, technical proofs, and additional simulations.

2 Causal graphical model with confounders

2.1 Structural equations with confounders

We consider a structural equation model with p primary variables $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ and q intervention variables $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$,

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

where $\mathbf{U}_{p\times p}$ is a matrix describing the causal influences among $\boldsymbol{Y}$, $\mathbf{W}_{q\times p}$ is a matrix representing the interventional effects of $\boldsymbol{X}$ on $\boldsymbol{Y}$, and $\boldsymbol{\varepsilon}$ is a vector of possibly correlated errors. Specifically,

- The parameter matrix $\mathbf{U}$, which is of primary interest, has a causal interpretation in that $\mathrm{U}_{kj} \neq 0$ indicates that Y_k is a cause of Y_j, denoted by $Y_k \to Y_j$. Thus, $\mathbf{U}$ represents a directed graph among primary variables. In what follows, we will focus on a directed acyclic graph (DAG), where no directed cycle is permissible and $\mathbf{U}$ is subject to the acyclicity constraint (Zheng et al., 2018; Yuan et al., 2019).
- The intervention variables $\boldsymbol{X}$ and errors $\boldsymbol{\varepsilon}$ are uncorrelated by reparameterization. As a result, $\mathbf{W}$ is associational instead of causal. Here, $\mathrm{W}_{lj} \neq 0$ indicates that X_l intervenes on Y_j, denoted by $X_l \to Y_j$. As $\boldsymbol{X}$ represents external interventions, no directed edge from a primary variable to an intervention variable is allowed.

- A non-diagonal $\boldsymbol{\Sigma}$ indicates the presence of unmeasured confounders. For instance, $\boldsymbol{\varepsilon} = \boldsymbol{\Phi}^\top \boldsymbol{\eta} + \boldsymbol{e}$ can be (not uniquely) written as a sum of correlated components $\boldsymbol{\Phi}^\top \boldsymbol{\eta}$ and independent components $\boldsymbol{e}$ so that $\boldsymbol{\Sigma} = \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$, where $\boldsymbol{\Phi}_{r \times p}$ is the matrix of confounding effects, $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{I}_{r \times r})$ represents r independent confounding sources, and $\boldsymbol{e} \sim N(\mathbf{0}, \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2))$ represents p independent errors. Whenever $\Sigma_{jk} \neq 0$ for some distinct (j, k), we have $\Sigma_{jk} = \sum_{m=1}^{r} \Phi_{mj} \Phi_{mk} \neq 0$, implying that some confounding variable η_m influences both Y_j and Y_k.

As such, $(\mathbf{U}, \mathbf{W})$ together represents a directed graph of p primary variables and q intervention variables, denoted as $\mathcal{G} = (\boldsymbol{X}, \boldsymbol{Y}, \mathcal{E}, \mathcal{I})$, where $\mathcal{E} = \{(k, j) : \mathrm{U}_{kj} \neq 0\}$ is the set of primary variable edges and $\mathcal{I} = \{(l, j) : \mathrm{W}_{lj} \neq 0\}$ is the set of intervention edges. In $\mathcal{G}$, (a) if $Y_k \to Y_j$, then Y_k is a parent of Y_j, and Y_j is a child of Y_k, (b) if $Y_k \to \cdots \to Y_j$ (a directed path from Y_k to Y_j), then Y_k is an ancestor of Y_j, and Y_j is a descendant of Y_k, and (c) if $Y_k \to \cdots \to Y_m \to \cdots \to Y_j$, then Y_m is a mediator of Y_k and Y_j. In what follows, for a graph $\mathcal{G}$, denote the parent set of Y_j as $\mathrm{pa}_{\mathcal{G}}(j) = \{k : Y_k \to Y_j\}$, the ancestor set of Y_j as $\mathrm{an}_{\mathcal{G}}(j) = \{k : Y_k \to \cdots \to Y_j\}$, and the intervention set of Y_j as $\mathrm{in}_{\mathcal{G}}(j) = \{l : X_l \to Y_j\}$. For ($k$, j) such that $Y_k \to \cdots \to Y_j$, denote the mediator set as $\mathrm{me}_{\mathcal{G}}(k, j) = \{m : Y_k \to \cdots \to Y_m \to \cdots \to Y_j\}$.

2.2 Identifiability and instrumental variables

The causal parameter matrix $\mathbf{U}$ is generally non-identifiable[1] without further conditions on the Gaussian errors $\boldsymbol{\varepsilon}$ or the interventions $\boldsymbol{X}$. Without invoking external interventions ($\mathbf{W} \equiv \mathbf{0}$), $\mathbf{U}$ can be identified under a certain error-scale assumption (Peters and Bühlmann, 2014; Ghoshal and Honorio, 2018; Rajendran et al., 2021), which is sensitive to variable scaling such as the common practice of standardizing variables (Reisach et al., 2021). To overcome this limitation, interventions are introduced to identify the causal parameters. With suitable interventions, $\mathbf{U}$ is identifiable if no confounder is present in the model ($\boldsymbol{\Sigma}$ is diagonal) (Oates et al., 2016; Chen et al., 2018; Li et al., 2023a). In addition, it is worth mentioning that $\mathbf{U}$ can be estimated without intervention if the errors $\boldsymbol{\varepsilon}$ are non-Gaussian

(Shimizu et al., 2006; Zhao et al., 2022); however, such methods are not applicable in the case of unmeasured confounding.

This subsection establishes the identifiability of (1) in the presence of unmeasured confounders using uncertain additive interventions (the targets of interventions are unknown) as IVs. To proceed, we introduce the notion of IV for our purpose.

Definition 1. An intervention variable X_l is said to be a valid IV of Y_k in $\mathcal{G}$ if **(IV1)** X_l intervenes on Y_k, namely $\mathrm{W}_{lk} \neq 0$, and **(IV2)** X_l does not intervene on any other primary variable $Y_{k'}$, namely $\mathrm{W}_{lk'} = 0$ for $k' \neq k$. Otherwise, X_l is called an invalid IV. Denote the valid IV set of Y_k as $\mathrm{iv}_{\mathcal{G}}(k) = \{l : X_l \to Y_k, X_l \not\to Y_{k'}, k' \neq k\}$.

Remark 1. Consider a bivariate case where we are interested in the potential causal effect $Y_1 \to Y_2$. In causal inference literature (Angrist et al., 1996; Kang et al., 2016), a valid IV X of Y_1 is required to satisfy that (a) X is related to the Y_1, referred to as relevance, (b) X has no directed edge to Y_2, called exclusion, and (c) X is not related to unmeasured confounders, called unconfoundedness. In (1), (IV1) is indeed the relevance property, (IV2) generalizes the exclusion property for causal discovery, and the requirement $\mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = \mathbf{0}$ corresponds to the unconfoundedness.

To identify $\mathbf{U}$, two challenges emerge as the confounders arise. First, determining causal directions in the graph becomes more challenging. In (1), because of hidden confounding, the distribution $\mathbb{P}_{(\boldsymbol{Y} \mid \boldsymbol{X})}$ does not admit the causal Markov property (Pearl, 2009) according to $\mathcal{G}$, that is, Y_j is not independent of its non-descendants given $(\boldsymbol{Y}_{\textsc{pa}_{\mathcal{G}}(j)}, \boldsymbol{X})$. As a result, the existing methods based on this property can learn wrong causal directions due to misspecification. To identify causal directions, we formalize the concept of unmediated parents to highlight the causal relations that are critical in identification.

Definition 2. A primary variable Y_k is an unmediated parent of Y_j in $\mathcal{G}$ if $Y_k \to Y_j$ and there is no other directed path from Y_k to Y_j. In other words, Y_k is an unmediated parent of Y_j if no mediator is between Y_k and Y_j.

Another challenge comes from uncertain interventions and invalid IVs. Assigning valid IVs for each primary variable can be difficult when the targets of interventions

are unknown. Thus, it may be effective to construct a set of candidate IVs (including invalid IVs) for each primary variable, on which we estimate the causal parameters $\mathbf{U}$. To this end, we define p candidate IV sets, one for each primary variable.

Definition 3. An intervention variable X_l is said to be a candidate IV of Y_k in $\mathcal{G}$ if **(IV1')** X_l intervenes on Y_k, and **(IV2')** X_l does not intervene on any non-descendant of Y_k. Denote the candidate IV set of Y_k by $\mathrm{cand}_{\mathcal{G}}(k) = \{l : X_l \to Y_k, X_l \to Y_j \text{ only if } k \in \mathrm{an}_{\mathcal{G}}(j)\}$.

The candidate IVs of Y_k include all valid IVs of Y_k, but not vice versa. A candidate IV of Y_k may be invalid, as it could intervene on descendants of Y_k.

Theorem 1 (Identifiability). *Suppose*

- *(A1)* $\mathrm{Cov}(\boldsymbol{X})$ *is positive definite.*
- *(A2)* $\mathrm{Cov}(Y_j, X_l \mid \boldsymbol{X}_{\{1,\dots,q\}\setminus\{l\}}) \neq 0$ *whenever* X_l *intervenes on an unmediated parent of* Y_j.
- *(A3) (Majority rule)* $|\mathrm{iv}_{\mathcal{G}}(k)| > |\mathrm{cand}_{\mathcal{G}}(k)|/2$; $k = 1, \dots, p$.

Then $(\mathbf{U}, \mathbf{W}, \boldsymbol{\Sigma})$ *in (1) are identifiable in that if* $(\mathbf{U}, \mathbf{W}, \boldsymbol{\Sigma})$ *and* $(\mathbf{U}', \mathbf{W}', \boldsymbol{\Sigma}')$ *encode the same probability distribution, then* $(\mathbf{U}, \mathbf{W}, \boldsymbol{\Sigma}) = (\mathbf{U}', \mathbf{W}', \boldsymbol{\Sigma}')$.

To our knowledge, Theorem 1 is a new result for Gaussian DAG with hidden confounding, establishing the identifiability of all parameters in (1). In fact, if the causal parameter $\mathbf{U}$ is identifiable, then so are parameters $\mathbf{W}, \boldsymbol{\Sigma}$. Regarding the conditions, (A1) states that $\mathrm{Cov}(\boldsymbol{X})$ has full rank, which is common in the IV literature (Kang et al., 2016; Chen et al., 2018). Note that (A1) permits discrete IV variables such as SNPs in data analysis. (A2) requires the interventional effects through unmediated parents not to cancel out when an invalid IV has multiple targets. (A3) requires valid IVs to dominate invalid ones so that the causal effect can be identified in the presence of latent confounders. Such a condition has been used in the causal inference literature (Kang et al., 2016; Windmeijer et al., 2019). As shown in Supplementary Materials Section 1, when (A3) fails, (1) can be non-identifiable. By comparison, (A1)–(A2) together with (A4) are used for model identification in the absence of unmeasured confounding (Li et al., 2023a).

- Each Y_k is intervened by at least one valid IV.

Noting that (A4) is implied by (A3), treating hidden confounding demands stronger conditions in view of Theorem 1.

3 Causal discovery

This section proposes a novel IV method to learn a DAG with unmeasured confounders. First, we introduce the ancestral relation graph (ARG), which, together with the candidate IV sets in Section 2.2, constitutes a basis for the proposed method.

Definition 4 (Ancestral relation graph). For a DAG $\mathcal{G} = (\boldsymbol{X}, \boldsymbol{Y}; \mathcal{E}, \mathcal{I})$, its ancestral relation graph is defined as $\mathcal{G}^+ = (\boldsymbol{X}, \boldsymbol{Y}; \mathcal{E}^+, \mathcal{I}^+)$, where

$$\mathcal{E}^+ = \{(k, j) : k \in \mathrm{an}_{\mathcal{G}}(j)\}, \qquad \mathcal{I}^+ = \{(l, j) : l \in \bigcup_{k \in \mathrm{an}_{\mathcal{G}}(j) \cup \{j\}} \mathrm{in}_{\mathcal{G}}(k)\}.$$

Here, $\mathcal{G}^+$ is a super-DAG of $\mathcal{G}$ in that $\mathcal{E}^+ \supseteq \mathcal{E}$ is the set of ancestral relations, $\mathcal{I}^+ \supseteq \mathcal{I}$ is a superset of interventional relations, and $\mathcal{G}^+$ is acyclic. Note that $\mathcal{E}^+$ defines a partial order for the primary variables $\boldsymbol{Y}$ in that $Y_k \prec_{\mathcal{G}} Y_j$ whenever $(k, j) \in \mathcal{E}^+$. Without confounding, $\mathbf{U}$ can be consistently estimated via direct regressions according to the known $\mathcal{G}^+$ (Shojaie and Michailidis, 2010), where $\mathcal{G}^+$ can be recovered by the peeling algorithm (Li et al., 2023a). However, this approach no longer applies in the presence of hidden confounders.

To address this obstacle, Sections 3.1–3.2 modify the peeling algorithm to construct the ARG $\mathcal{G}^+$ and the candidate IV sets $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1 \le k \le p}$, and then Sections 3.3–3.4 develop a method to estimate $\mathbf{U}$ assuming the ARG and candidate IVs are known.

3.1 Identification of $\mathcal{G}^+$ and candidate IVs

In this subsection, we modify the peeling algorithm, originally designed for a model without unmeasured confounders (Li et al., 2023a), to uncover $\mathcal{G}^+$ and $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1 \le k \le p}$ in the presence of hidden confounders, of which the results can be subsequently used as the inputs for identification of $\mathbf{U}$ in Section 3.3. The modified peeling

algorithm essentially requires p regressions to identify the ARG and candidate IVs, which is suited for large-scale causal discovery. Moreover, the produced ARG and candidate IV sets enjoy desirable statistical properties; see Section 5.

Let us begin with an observation that (1) can be rewritten as

$$\boldsymbol{Y} = \mathbf{V}^{\top}\boldsymbol{X} + (\mathbf{I} - \mathbf{U}^{\top})^{-1}\boldsymbol{\varepsilon}, \qquad (2)$$

where $\mathbf{V} = \mathbf{W}(\mathbf{I} - \mathbf{U})^{-1}$ and $\mathrm{V}_{lj} = \sum_{k=1}^{p} \mathrm{W}_{lk}(\mathrm{I}_{kj} + \mathrm{U}_{kj} + \cdots + (\mathbf{U}^{p-1})_{kj})$. Intuitively, $\mathrm{V}_{lj} \neq 0$ implies the dependence of Y_j on X_l through a directed path $X_l \to Y_k \to \cdots \to Y_j$, and hence that X_l intervenes on Y_j itself (when $k = j$) or its ancestor Y_k (when $k \neq j$). In cases where X_l intervenes exclusively on one primary variable, the following proposition provides insights into the connection between $\mathbf{V}$ and $\mathcal{G}^{+}$.

Proposition 1. *Suppose Assumptions (A1), (A2), and (A4) are satisfied. There exists at least one intervention variable X_l such that $\mathrm{V}_{lk} \neq 0$ and $\mathrm{V}_{lk'} = 0$ for $k' \neq k$ if and only if Y_k is a leaf node (has no descendant). Moreover, such X_l is a valid IV of Y_k in $\mathcal{G}$.*

Proposition 1 suggests that the leaves and their valid IVs in $\mathcal{G}$ can be identified by

$$\begin{aligned}
\mathrm{leaf}(\mathcal{G}) &= \{k : \text{for some } l, \mathrm{V}_{lk} \neq 0 \text{ and } \mathrm{V}_{lk'} = 0 \text{ for all } k' \neq k\} \\
&= \{k : k = \arg\max_{j} |\mathrm{V}_{lj}| \text{ for some } l = \arg\min_{|\mathbf{v}_{l,+}|_0 > 0} \|\mathbf{V}_{l,+}\|_0\}, \\
\mathrm{iv}_{\mathcal{G}}(k) &= \{l : \mathrm{V}_{lk} \neq 0 \text{ and } \mathrm{V}_{lk'} = 0 \text{ for all } k' \neq k\} \\
&= \{l : l = \arg\min_{|\mathbf{v}_{l,+}|_0 > 0} \|\mathbf{V}_{l,+}\|_0 \text{ and } k = \arg\max_{j} |\mathrm{V}_{lj}|\}, \quad k \in \mathrm{leaf}(\mathcal{G}).
\end{aligned} \qquad (3)$$

After the leaf nodes are learned, we can remove them to obtain a sub-DAG. If X_l is a valid IV of a non-leaf Y_k in $\mathcal{G}$, its validity for Y_k is retained in the sub-DAG, implying (A4) continues to hold. Moreover, Assumptions (A1)–(A2) are naturally upheld in the sub-DAG. Hence, the requirements of Proposition 1 are satisfied in the sub-DAG, whose leaf variables and their valid IVs can be learned in the same fashion. As a result, we can successively identify and remove (i.e., peel) the leaf nodes from the DAG and sub-DAGs. This yields a topological order of primary variables but does not recover $\mathcal{G}^{+}$.

Next, we investigate how $\mathbf{V}$ can be further used to recover $\mathcal{G}^+$ with $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1\le k\le p}$. Subsequently, we use $\mathcal{G}^- = (\boldsymbol{X}^-, \boldsymbol{Y}^-; \mathcal{E}^-, \mathcal{I}^-)$ to denote a generic sub-DAG produced by peeling, where $\boldsymbol{Y}^-$ are the primary variables in $\mathcal{G}^-$ and $\boldsymbol{Y} \setminus \boldsymbol{Y}^-$ are peeled ones, $\boldsymbol{X}^-$ are intervention variables on $\boldsymbol{Y}^-$, $\mathcal{E}^-$ is the set of causal relations among $\boldsymbol{Y}^-$, and $\mathcal{I}^-$ is the set of interventional relations between $\boldsymbol{X}^-$ and $\boldsymbol{Y}^-$. Then each variable in $\boldsymbol{Y}^-$ is a non-descendant of each in $\boldsymbol{Y} \setminus \boldsymbol{Y}^-$. Moreover, $\mathrm{leaf}(\mathcal{G}^-)$ and $\{\mathrm{iv}_{\mathcal{G}^-}(k)\}_{k\in\mathrm{leaf}(\mathcal{G}^-)}$ are identified by (3).

Proposition 2. *Suppose Assumptions (A1), (A2), and (A4) are satisfied. Let Y_k be a leaf node in $\mathcal{G}^-$ and Y_j be in $\boldsymbol{Y} \setminus \boldsymbol{Y}^-$. Then the following statements are true.*

(A) If $\mathrm{V}_{lj} \neq 0$ *for all* $l \in \mathrm{iv}_{\mathcal{G}^-}(k)$*, we have* $(k, j) \in \mathcal{E}^+$.

(B) If Y_k is an unmediated parent of Y_j, then $\mathrm{V}_{lj} \neq 0$ *for all* $l \in \mathrm{iv}_{\mathcal{G}^-}(k)$.

Proposition 2 outlines a method for identifying edges in $\mathcal{G}^+$ from the leaf variables of $\mathcal{G}^-$ to the peeled variables $\boldsymbol{Y} \setminus \boldsymbol{Y}^-$ by

$$\{(k, j) : Y_k \in \mathrm{leaf}(\mathcal{G}^-), Y_j \in \boldsymbol{Y} \setminus \boldsymbol{Y}^- \text{ and } \mathrm{V}_{lj} \neq 0 \text{ for all } l \in \mathrm{iv}_{\mathcal{G}^-}(k)\}. \quad (4)$$

Specifically, (A) shows that any identified edge must be present in $\mathcal{G}^+$, so no extra edges are identified. Meanwhile, (B) shows that every directed edge from an unmediated parent must be correctly discovered. Importantly, the collection of all such edges suffices to recover all ancestral relationships, which guarantees that no edge in $\mathcal{E}^+$ is overlooked. Upon the identification of $\mathcal{G}^+$, the candidate IV sets can be learned by

$$\mathrm{ca}_{\mathcal{G}}(k) = \{l : (l, k) \in \mathcal{I}^+ \text{ and } (l, j) \in \mathcal{I}^+, k \neq j \text{ only if } (k, j) \in \mathcal{E}^+\}, \quad 1 \le k \le p. \quad (5)$$

Consequently, Propositions 1–2 enable the recovery of $\mathcal{G}^+$ and $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1\le k\le p}$.

3.2 Finite-sample estimation of $\mathcal{G}^+$ and candidate IVs

This subsection implements the modified peeling algorithm delineated in Section 3.1 to estimate $\mathcal{G}^+$ and $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1\le k\le p}$. To proceed, suppose data matrices $\mathbf{Y}_{p\times n} = (\mathbf{Y}_{+,1}, \ldots, \mathbf{Y}_{+,n})$ and $\mathbf{X}_{q\times n} = (\mathbf{X}_{+,1}, \ldots, \mathbf{X}_{+,n})$ are given, where $(\mathbf{Y}_{+,i}, \mathbf{X}_{+,i})_{i=1}^n$ are

sampled from (1) independently. We estimate $\mathbf{V}$ by $\mathbf{V} = (\mathbf{V}_{+,1}, \ldots, \mathbf{V}_{+,p})$ with sparse regressions

$$\mathbf{V}_{+,j} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (\mathrm{Y}_{j,i} - \boldsymbol{\beta}^{\top} \mathbf{X}_{+,i})^2 \quad \text{s.t.} \quad \| \boldsymbol{\beta} \|_0 \leq \kappa'_j \quad (6)$$

where $1 \leq \kappa'_j \leq q$ is tuned by BIC for $1 \leq j \leq p$ Moreover, the truncated Lasso penalty (TLP) (Shen et al., 2012) is used as the computational surrogate for $\|\cdot\|_0$, where TLP is defined as $\mathrm{TLP}_{\tau}(\boldsymbol{\beta}) = \sum_{j=1}^{r} \min(|\beta_j| / \tau, 1)$ for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)$, and $\tau > 0$ is a hyperparameter in TLP; see Supplementary Materials Section 2 for details. The modified peeling algorithm based on Section 3.1 is summarized in Algorithm 1.[2]

Algorithm 1: Estimation of $\mathcal{G}^+$ and $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1 \leq k \leq p}$

Input: Data $\mathbf{Y}_{p \times n}$ and $\mathbf{X}_{q \times n}$;

1 Compute $\mathbf{V}$ via (6);

2 Initialize $\mathbf{V} \leftarrow \mathbf{V}$, $\mathcal{E}^+ \leftarrow \varnothing$, $\mathcal{I}^+ \leftarrow \{(l,k) : \mathrm{V}_{lk} \neq 0\}$;

3 Initialize $\mathcal{G}^-$ by $\boldsymbol{Y}^- \leftarrow \boldsymbol{Y}$, $\boldsymbol{X}^- \leftarrow \boldsymbol{X}$, $\mathcal{E}^- \leftarrow \mathcal{E}^+$, $\mathcal{I}^- \leftarrow \mathcal{I}^+$;

4 while $\boldsymbol{Y}^-$ *is not empty* **do**

5 Update $\mathrm{leaf}(\mathcal{G}^-)$ and $\{\mathrm{iv}_{\mathcal{G}^-}(k)\}_{k \in \mathrm{leaf}(\mathcal{G}^-)}$ via (3);

6 Update $\mathcal{E}^+$ by adding (4);

7 Update $\mathcal{G}^-$ by removing $\mathrm{leaf}(\mathcal{G}^-)$ and $\mathbf{V}$ by keeping the columns in $\boldsymbol{Y}^-$;

8 end

9 Update $\mathcal{E}^+ \leftarrow \{(k,j) : Y_k \rightarrow \cdots \rightarrow Y_j \text{ in } \mathcal{E}^+\}$;

10 Update $\mathcal{I}^+ \leftarrow \{(l,j) : (l,k) \in \mathcal{I}^+ \text{ and } (k,j) \in \mathcal{E}^+\}$;

11 Update $\mathrm{ca}_{\mathcal{G}}(k)$ by (5);

12 return $\mathcal{E}^+$, $\mathcal{I}^+$, *and* $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1 \leq k \leq p}$;

3.3 Identification of $\mathbf{U}$

In this subsection, we present a new method for identifying causal effects $\mathbf{U}$, using the ARG $\mathcal{G}^+$ and candidate IV sets $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1 \leq k \leq p}$ as inputs. Note that $\{\mathrm{an}_{\mathcal{G}}(k)\}_{1 \leq k \leq p}$

and $\{\mathrm{me}_{\mathcal{G}}(k,j)\}_{(k,j)\in\mathcal{E}^+}$ can be derived from $\mathcal{G}^+$. Throughout this subsection, the subscript $\mathcal{G}$ is dropped for brevity and α, β, γ denote nuisance parameters in regression. Moreover, we assume that ε and $\boldsymbol{X}$ are independent to simplify the derivation; see Lemmas 1–2 in the Appendix for the case with ε and $\boldsymbol{X}$ being uncorrelated.

The case with all IVs being valid.

We begin with a special case of (1) where all IVs are valid, that is, $\mathrm{ca}(k) = \mathrm{iv}(k)$; $k = 1, \ldots, p$.

To estimate $\mathbf{U}$, note that $\mathbf{U}$ is supported on $\mathcal{E}^+$, namely $\mathbf{U} = (\mathbf{U}_{\mathcal{E}^+}, \mathbf{0})$. Here, we consider estimating U_{kj}, as well as selecting nonzero U_{kj} for graph recovery, for each $(k,j) \in \mathcal{E}^+$, as described in Figure 1 (a).

To pinpoint the difficulties and motivate our approach, we make the following observations. First, regression of Y_j on Y_k together with covariates $(\boldsymbol{Y}_{\mathrm{an}(j)\setminus\{k\}}, \boldsymbol{X})$ can bias the estimation due to confounder η. Second, in hope of treating confounders one might replace Y_k with its surrogate $\mathbb{E}(Y_k \mid \boldsymbol{Y}_{\mathrm{an}(k)}, \boldsymbol{X})$ to regress Y_j on $\mathbb{E}(Y_k \mid \boldsymbol{Y}_{\mathrm{an}(k)}, \boldsymbol{X})$ with $(\boldsymbol{Y}_{\mathrm{an}(j)\setminus\{k\}}, \boldsymbol{X}_{\mathrm{iv}(k)^c})$ being covariates. However, this is also problematic. For explanation, note that $\mathrm{an}(j)\setminus\{k\}$ can be partitioned into mediators $\mathrm{me}(k,j)$ and non-mediators

$$\mathrm{nm}(k,j) = \mathrm{an}(j) \setminus (\mathrm{me}(k,j) \cup \{k\}).$$

In Figure 1 (a), $\boldsymbol{X}_{\mathrm{iv}(k)}$ can be associated with η given $\boldsymbol{Y}_{\mathrm{an}(j)\setminus\{k\}} = (\boldsymbol{Y}_{\mathrm{me}(k,j)}, \boldsymbol{Y}_{\mathrm{nm}(k,j)})$, violating the unconfoundedness of IVs (Remark 1) and causing an estimation bias. This is because the mediators $\boldsymbol{Y}_{\mathrm{me}(k,j)}$ generate additional associations after conditioning on them; see the Appendix for technical discussion using the concept of d-separation (Pearl, 2009).

Now, we propose a new method, which eliminates the impact of mediators $\boldsymbol{Y}_{\mathrm{me}(k,j)}$ by introducing the working response $\bar{Y}_j = Y_j - \mathbf{U}^{\top}_{\mathrm{me}(k,j),j}\boldsymbol{Y}_{\mathrm{me}(k,j)}$, as depicted in Figure 1 (b). Of note, the definition of $\bar{Y}_j$ depends on (k, j), which is dropped for simplicity. As in Angrist et al. (1996), we have

$$
\begin{aligned}
& \mathbb{E}\left(\bar{Y}_j \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}\right) \\
\overset{\text{(i)}}{=} \; & \mathrm{U}_{kj}\,\mathbb{E}\left(Y_k \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}\right) + \sum_{k' \in \mathrm{nm}(k,j)} \mathrm{U}_{k'j} Y_{k'} + \sum_{l \notin \mathrm{iv}(k)} \mathrm{W}_{lj} X_l + \mathbb{E}\left(\varepsilon_j \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}\right) \qquad (7) \\
\overset{\text{(ii)}}{=} \; & \mathrm{U}_{kj} \tilde{Y}_k + \boldsymbol{\gamma}^\top \boldsymbol{Z},
\end{aligned}
$$

where $\tilde{Y}_k = \mathbb{E}(Y_k \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X})$, $\boldsymbol{Z} = (\boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}_{\mathrm{ca}(k)^c}) = (\boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}_{\mathrm{iv}(k)^c})$, equality (i) follows from (1), and equality (ii) holds because $\mathbb{E}\left(\varepsilon_j \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}\right)$ is a linear combination of $(\boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}_{\mathrm{iv}(k)^c})$ by Lemma 1 in Appendix. Observe that $\tilde{Y}_k$ depends on $\boldsymbol{X}_{\mathrm{iv}(k)}$ while $\boldsymbol{Z}$ does not. As a result, the U_{kj} is identified through the working response regression.

This approach requires the knowledge of $\mathbf{U}_{\mathrm{me}(k,j),j}$ prior to identifying U_{kj}. Given $\mathcal{G}^+$, we develop a sequential procedure to learn **U**. First, we identify U_{kj} for each pair (*k*, *j*) such that the longest path in $\mathcal{G}^+$ between *k* and *j* is equal to *d* = 1. Then for (*k*, *j*) such that the longest path in $\mathcal{G}^+$ between *k* and *j* is *d* = 2, the effects of mediators $\mathbf{U}_{\mathrm{me}(k,j),j}$ are available. Thus, we can identify U_{kj} in (7). Proceed similarly for $d = 3, 4, 5, \ldots$ until all pairs in $\mathcal{E}^+$ have been identified.

The case with invalid IVs.

In general, $\mathrm{ca}(k) \supseteq \mathrm{iv}(k)$ because of invalid IVs, where $\mathrm{ca}(k)$ is known but $\mathrm{iv}(k)$ is unknown. Similar to Kang et al. (2016), we have

$$
\begin{aligned}
& \mathbb{E}\left(\bar{Y}_j \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}\right) \\
= \; & \mathrm{U}_{kj}\,\mathbb{E}\left(Y_k \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}\right) + \sum_{k' \in \mathrm{nm}(k,j)} \mathrm{U}_{k'j} Y_{k'} + \sum_{l \notin \mathrm{iv}(k)} \mathrm{W}_{lj} X_l + \mathbb{E}\left(\varepsilon_j \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}\right) \qquad (8) \\
\overset{\text{(iii)}}{=} \; & \mathrm{U}_{kj} \tilde{Y}_k + \boldsymbol{\gamma}^\top \boldsymbol{Z} + \sum_{l \in \mathrm{ca}(k) \setminus \mathrm{iv}(k)} \beta_l X_l,
\end{aligned}
$$

where $\tilde{Y}_k = \mathbb{E}(Y_k \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X})$, $\boldsymbol{Z} = (\boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}_{\mathrm{ca}(k)^c})$, equality (iii) holds by Lemma 1 in Appendix, and $\beta_l = \mathrm{W}_{lj} \neq 0$ indicates X_l is an invalid IV for Y_k. However, since $\mathrm{iv}(k)$ has not been identified and $\tilde{Y}_k$ depends on $\boldsymbol{X}_{\mathrm{ca}(k)}$, the representation of (iii) may not be unique. When the majority rule (A3) is satisfied by the DAG, the term (iii) admits the unique expression as in (8), providing the identification of U_{kj}. This leads to a sparse regression for an infinite sample

$$\min_{\mathrm{U}_{kj},\boldsymbol{\beta},\boldsymbol{\gamma}} \mathbb{E}\left(\overline{Y}_j - \mathrm{U}_{kj}\tilde{Y}_k - \boldsymbol{\gamma}^\top \boldsymbol{Z} - \boldsymbol{\beta}^\top \boldsymbol{X}_{\mathrm{ca}(k)}\right)^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \le \kappa, \tag{9}$$

where $0 \le \kappa < |\mathrm{ca}(k)|/2$ is an integer-valued hyperparameter controlling the sparsity of $\boldsymbol{\beta}$.

3.4 Finite-sample estimation of U

Suppose $(\mathbf{Y}_{p\times n}, \mathbf{X}_{q\times n})$ are given. To estimate U_{kj}, noting that $\tilde{Y}_k$ is linear in $(\boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X})$ by Lemma 1, we estimate $\mathrm{Y}_{k,i}$ by $\mathrm{Y}_{k,i} = \boldsymbol{\alpha}_1^\top \mathbf{X}_{+,i} + \boldsymbol{\alpha}_2^\top \mathbf{Y}_{\mathrm{nm}(k),i}$, where $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ solves

$$\min_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2} \sum_{i=1}^{n} \left(\mathrm{Y}_{k,i} - \boldsymbol{\alpha}_1^\top \mathbf{X}_{+,i} + \boldsymbol{\alpha}_2^\top \mathbf{Y}_{\mathrm{nm}(k),i}\right)^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}_1\|_0 + \|\boldsymbol{\alpha}_2\|_0 \le \nu_1, \tag{10}$$

with ν_1 being a tuning parameter. Let the final estimate U_{kj} with $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ be the solution to the working response regression (provided that $\mathbf{U}_{\mathrm{me}(k,j),j}$ are available)

$$\begin{aligned}\min_{\mathrm{U}_{kj},\boldsymbol{\beta},\boldsymbol{\gamma}} & \sum_{i=1}^{n} \left(\left(\mathrm{Y}_{j,i} - \mathbf{U}_{\mathrm{me}(k,j),j}^\top \mathbf{Y}_{\mathrm{me}(k,j),i}\right) - \mathrm{U}_{kj}\mathrm{Y}_{k,i} - \boldsymbol{\beta}^\top \boldsymbol{X}_{\mathrm{ca}(k),i} - \boldsymbol{\gamma}^\top \boldsymbol{Z}_i\right)^2 \\ \text{s.t.} \quad & \rho(\mathrm{U}_{\mathrm{kj}}) + \|\boldsymbol{\beta}\|_0 \le \kappa, \quad \|\boldsymbol{\gamma}\|_0 \le \nu_2,\end{aligned} \tag{11}$$

where $0 \le \kappa \le |\mathrm{ca}(k)|/2$ and $0 \le \nu_2 \le |\mathrm{nm}(k,j)| + |\mathrm{ca}(k)^c|$ are tuning parameters. Depending on the purpose, $\rho(\cdot) = \mathrm{I}(\cdot \ne 0)$ for graph recovery and $\rho(\cdot) = 0$ for effect estimation without selection. In (10)–(11), ν_1, ν_2 are added to treat possible high-dimensional situations and the hyperparameters are tuned by BIC. Algorithm 2 summarizes the procedure.

Algorithm 2: Estimation of $\mathbf{U}$

Input: Data $\mathbf{Y}_{p\times n}$ and $\mathbf{X}_{q\times n}$, ARG $\mathcal{G}^+$ and candidate IV sets $\{\mathrm{ca}_{\mathcal{G}}(k)\}_{1\le k\le p}$;

1 Initialize $\mathbf{U} \leftarrow \mathbf{0}$ and $d \leftarrow 1$;

2 while $d \le$ *the length of the longest directed path in* $\mathcal{G}^+$ **do**

3 For $(k,j) \in \mathcal{E}^+$ so that the length of the longest directed path from Y_k to Y_j is d, estimate U_{kj} with (10)–(11);

4 Update $d \leftarrow d + 1$;

5 end

6 return $\mathbf{U}$;

---

4 Likelihood inference

This section develops a likelihood ratio test for the presence of multiple directed edges. Let $\mathcal{H} \subseteq \{(k, j) : k \neq j, 1 \leq k, j \leq p\}$ be a hypothesized edge set for primary variables $\boldsymbol{Y}$, where $(k, j) \in \mathcal{H}$ specifies a (hypothesized) directed edge $Y_k \to Y_j$ in (1). Now consider simultaneous testing of directed edges,

$$H_0 : \mathrm{U}_{kj} = 0 \text{ for all } (k, j) \in \mathcal{H} \quad \text{versus} \quad H_a : \mathrm{U}_{kj} \neq 0 \text{ for some } (k, j) \in \mathcal{H}. \tag{12}$$

The null hypothesis H_0 asserts that all hypothesized edges in $\mathcal{H}$ are absent in the true graph $\mathcal{G}$. Rejecting H_0 indicates that at least one hypothesized edge in $\mathcal{H}$ presents in $\mathcal{G}$.

The likelihood ratio.

Given $\mathcal{G}^+ = (\boldsymbol{X}, \boldsymbol{Y}; \mathcal{E}^+, \mathcal{I}^+)$, let $\boldsymbol{\theta}(\mathcal{G}^+) = (\mathbf{U}, \mathbf{W})$ encode the coefficient parameters in $\mathcal{G}^+$, where $\mathbf{U} = (\mathbf{U}_{\mathcal{E}^+}, \mathbf{0})$ and $\mathbf{W} = (\mathbf{W}_{\mathcal{I}^+}, \mathbf{0})$. As such, the adjacency matrix $\mathbf{U}$ automatically meets the acyclicity constraint. Given a random sample $(\mathbf{Y}_{+,i}, \mathbf{X}_{+,i})_{i=1}^n$, the log-likelihood is written as (up to an additive constant)

$$L(\boldsymbol{\theta}(\mathcal{G}^+), \boldsymbol{\Omega}) = -\frac{1}{2} \sum_{i=1}^{n} \left\| \boldsymbol{\Omega}^{1/2} \left((\mathbf{I} - \mathbf{U}^\top) \mathbf{Y}_{+,i} - \mathbf{W}^\top \mathbf{X}_{+,i} \right) \right\|_2^2 + \frac{n}{2} \log\det(\boldsymbol{\Omega}), \tag{13}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the inverse of $\boldsymbol{\Sigma}$ in (1). Then the maximum likelihood estimation (MLE) of (1) can be written as

$$\max_{(\mathcal{G}^+, \boldsymbol{\Omega})} \max_{\boldsymbol{\theta}(\mathcal{G}^+)} L(\boldsymbol{\theta}(\mathcal{G}^+), \boldsymbol{\Omega}). \tag{14}$$

In view of (14), to obtain a likelihood ratio statistic for (12) we need to compute the following quantities: (1) a consistent estimate $\mathcal{G}^+$ of $\mathcal{G}^+$, (2) a consistent estimate $\boldsymbol{\Omega}$ of $\boldsymbol{\Omega}$, and (3) two estimates, $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\theta}^{(1)}$, of $\boldsymbol{\theta}(\mathcal{G}^+)$ under H_0 and H_a, respectively. This leads to the likelihood ratio defined as

$$L(\boldsymbol{\theta}^{(1)}, \boldsymbol{\Omega}) - L(\boldsymbol{\theta}^{(0)}, \boldsymbol{\Omega}), \qquad (15)$$

where $\mathcal{G}^+$ is estimated by Algorithm 1 and $\boldsymbol{\Omega}$ is estimated from the residuals after fitting model (1) via Algorithm 2.

4 Inference subject to acyclicity.

In classical models, a likelihood ratio of form (15) has a nondegenerate and tractable limiting distribution, typically a chi-squared distribution with degrees of freedom $|\mathcal{H}|$. However, the likelihood ratio for (12) may behave differently from classical ones since (15) may be degenerate or intractable, as to be explained.

First, note that the maximum likelihood subject to a wrong ARG $\mathcal{G}^+ \supsetneq \mathcal{G}$ tends to be smaller than that subject to the correct $\mathcal{G}^+$, that is,

$$\max_{\mathcal{G}^+ \supsetneq \mathcal{G}} \max_{\boldsymbol{\theta}(\mathcal{G}^+), \boldsymbol{\Omega}} L(\boldsymbol{\theta}(\mathcal{G}^+), \boldsymbol{\Omega}) < \max_{\boldsymbol{\theta}(\mathcal{G}^+), \boldsymbol{\Omega}} L(\boldsymbol{\theta}(\mathcal{G}^+), \boldsymbol{\Omega}),$$

as $n \to \infty$ under some regularity conditions for consistency. Thus, we assume $\mathcal{G}^+ = \mathcal{G}^+$ in this paragraph. Then $\boldsymbol{\theta}^{(0)}$ is the MLE subject to $\mathcal{G}^+$ and $\mathbf{U}_{\mathcal{H}} = \mathbf{0}$, which is equal to the MLE subject to the graph $\mathcal{G}_0^+ = (\boldsymbol{X}, \boldsymbol{Y}; \mathcal{E}^+ \setminus \mathcal{H}, \mathcal{I}^+)$. Meanwhile, to test whether any edge in $\mathcal{H}$ exists, $\boldsymbol{\theta}^{(1)}$ is the MLE subject to an augmented graph $\mathcal{G}_1^+ = (\boldsymbol{X}, \boldsymbol{Y}; \mathcal{E}^+ \cup \mathcal{H}, \mathcal{I}^+)$ with hypothesized edges being added, namely, $\mathbf{U}^{(1)} = (\mathbf{U}^{(1)}_{\mathcal{E}^+ \cup \mathcal{H}}, \mathbf{0})$ and $\mathbf{W}^{(1)} = (\mathbf{W}^{(1)}_{\mathcal{I}^+}, \mathbf{0})$. Of note, since $\mathcal{H}$ is pre-specified by the user, $\mathcal{G}_1^+$ is not necessarily acyclic, and thus, not all edges in $\mathcal{H}$ could present in $\mathbf{U}^{(1)}$. Furthermore, if a hypothesized edge (*k*, *j*) is present in $\mathbf{U}^{(1)}$, then $\{(k, j)\} \cup \mathcal{E}^+$ must have no directed cycle and (15) is strictly positive (nondegenerate). However, even if (15) does not degenerate to zero, its limiting distribution can be complicated when there exist multiple ways of augmenting $\mathcal{G}^+$ with the edges in $\mathcal{H}$ while maintaining the resulting graph as a DAG. Therefore, a regularity condition for $\mathcal{H}$ is necessary to rule out intractable situations.

On the ground of the foregoing discussion, we introduce the concepts of nondegeneracy and regularity to characterize the behavior of (15) as in Li et al. (2023a).

Definition 5 (Nondegeneracy and regularity with respect to [INEQ-START). $\mathcal{G}^+$]

(A) An edge $(k, j) \in \mathcal{H}$ is said to be nondegenerate with respect to an ancestral graph $\mathcal{G}^+ = (\boldsymbol{Y}, \boldsymbol{X}; \mathcal{E}^+, \mathcal{I}^+)$ if $\{(k, j)\} \cup \mathcal{E}^+$ contains no directed cycle. Otherwise, (*k*, *j*) is said to be 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 said to be nondegenerate with respect to $\mathcal{G}^+$ if $\mathcal{D} \neq \varnothing$. Otherwise, H_0 is said to be degenerate.

(B) A null hypothesis H_0 is said to be regular with respect to $\mathcal{G}^+$ if $\mathcal{D} \cup \mathcal{E}^+$ contains no directed cycle. Otherwise, H_0 is called irregular.

Suppose H_0 is nondegenerate and regular. Then $\theta^{(0)}$ is the MLE subject to the graph $\mathcal{G}_0^+ = (\boldsymbol{X}, \boldsymbol{Y}; \mathcal{E}^+ \setminus \mathcal{D}, \mathcal{I}^+)$ and $\theta^{(1)}$ is the MLE subject to the graph $\mathcal{G}_1^+ = (\boldsymbol{X}, \boldsymbol{Y}; \mathcal{E}^+ \cup \mathcal{D}, \mathcal{I}^+)$.

Now, we investigate the limiting distribution of (15) and derive an asymptotic test based on it. To this end, define the statistic

$$T(\mathcal{D}) = \begin{cases} 2\left(L(\theta^{(1)}, \boldsymbol{\Omega}) - L(\theta^{(0)}, \boldsymbol{\Omega})\right) & \text{if } |\mathcal{D}| \text{ is fixed,} \\ \left(2\left(L(\theta^{(1)}, \boldsymbol{\Omega}) - L(\theta^{(0)}, \boldsymbol{\Omega})\right) - |\mathcal{D}|\right) / \sqrt{2|\mathcal{D}|} & \text{if } |\mathcal{D}| \to \infty. \end{cases} \tag{16}$$

Theorem 2 (Limiting distribution). *Assume the null hypothesis H_0 is nondegenerate and regular. Suppose $\mathbb{P}(\mathcal{G}^+ = \mathcal{G}^+) \to 1$ as $n \to \infty$. Then we have $\mathbb{P}(\mathcal{D} = \mathcal{D}) \to 1$. In addition, if $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}\|_2^2 = O_{\mathbb{P}}(|S| \log(p \vee n)/n)$ where $S = \{(k, j) : \Omega_{kj} \neq 0\}$, then under H_0,*

$$T(\mathcal{D}) \xrightarrow{d} \begin{cases} \chi^2_{|\mathcal{D}|}, & \text{if } |\mathcal{D}| \text{ is fixed and } |S| \log(p \vee n)/n \to 0, \\ N(0,1), & \text{if } |\mathcal{D}| \to \infty \text{ and } |\mathcal{D}||S| \log(p \vee n)/n \to 0. \end{cases}$$

On the basis of Theorem 2, we conduct inference by substituting $|\mathcal{D}|$ by its estimate $|\mathcal{D}|$ and proceed with the empirical rule: (1) use the chi-squared test when $|\mathcal{D}| < 50$, and (2) use the normal test when $|\mathcal{D}| \geq 50$.

Theorem 2 requires a good estimator $\boldsymbol{\Omega}$ of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ to account for the confounding effects, where $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{\varepsilon})$. To estimate $\boldsymbol{\Omega}$, let $\hat{\boldsymbol{\varepsilon}}_{+,i} = (\mathbf{I} - \mathbf{U})^\top \mathbf{Y}_{+,i} - \mathbf{W}^\top \mathbf{X}_{+,i}$; $i = 1, \ldots, n$ be the estimated residuals after fitting (1) with Algorithm 2. Here we use the

neighborhood selection method (Meinshausen and Bühlmann, 2006) with an additional refitting to obtain a positive definite estimate $\boldsymbol{\Omega}$. In Supplementary Materials, we include the computational details and show that this estimator satisfies $\| \boldsymbol{\Omega} - \boldsymbol{\Omega} \|_F^2 = O_{\mathbb{P}}(|S| \log(p \vee n)/n)$ so that Theorem 2 applies.

Remark 2. In Theorem 2, we focus on nondegenerate and regular hypotheses. For a degenerate case, we define the p-value as one. For an irregular case where $\mathcal{D} \cup \mathcal{E}^+$ contains a directed cycle, we decompose H_0 into sub-hypotheses $H_0^{(1)}, \ldots, H_0^{(r)}$, each of which is regular. Then testing H_0 is reduced to multiple testing for $H_0^{(1)}, \ldots, H_0^{(r)}$.

Finally, we discuss two aspects of likelihood estimation and inference in the presence of unmeasured confounding. First, when $\boldsymbol{\Sigma}$ is non-diagonal, the likelihood in (13) cannot be factorized according to $\mathcal{G}$ (or $\mathcal{G}^+$). This implies that, unlike the case without latent confounders (Shojaie and Michailidis, 2010), the parameters of each equation in (1) cannot be estimated separately given $\mathcal{G}^+$. Indeed, the likelihood estimation of $(\mathbf{U}, \mathbf{W})$ in (1) requires a preliminary estimate of $\boldsymbol{\Omega}$ to account for correlations arising from hidden confounding. Furthermore, compared to Li et al. (2023a), the likelihood ratio (15) is no longer a sum of likelihood ratios of equations associated with nondegenerate hypothesized edges, rendering inference more challenging in both computation and theory when hidden confounders are present. Computationally, the likelihood ratio (15) requires maximization of the full likelihood, which is costly for a large-scale graph. Theoretically, estimating $\boldsymbol{\Omega}$ and $(\mathbf{U}, \mathbf{W})$ in high-dimensional situations may suffer from the curse of dimensionality.

Second, to mitigate the challenges in inference, we may conduct inference with respect to a sub-DAG to achieve dimensionality reduction. Specifically, let $\mathcal{D}$ be the nondegenerate edges of H_0. Given ARG $\mathcal{G}^+$, we perform likelihood inference using a sub-DAG (of ARG) $\mathcal{G}^+_{\text{sub}} = (\boldsymbol{X}_{\text{sub}}, \boldsymbol{Y}_{\text{sub}}; \mathcal{E}^+_{\text{sub}}, \mathcal{I}^+_{\text{sub}})$, where all edges specified in $\mathcal{D}$ are among primary variables $\boldsymbol{Y}_{\text{sub}}$, and $\boldsymbol{Y}_{\text{sub}}$ are non-descendants of $\boldsymbol{Y} \setminus \boldsymbol{Y}_{\text{sub}}$ in the graph $(\boldsymbol{X}, \boldsymbol{Y}; \mathcal{E}^+ \cup \mathcal{D}, \mathcal{I}^+)$, $\boldsymbol{X}_{\text{sub}}$ is the set of intervention variables of $\boldsymbol{Y}_{\text{sub}}$, $\mathcal{E}^+_{\text{sub}}$ is the set of ancestral relations among $\boldsymbol{Y}_{\text{sub}}$, and $\mathcal{I}^+_{\text{sub}}$ is the set of interventional relations between $\boldsymbol{X}_{\text{sub}}$ and $\boldsymbol{Y}_{\text{sub}}$ in ARG $\mathcal{G}^+$. Then the test statistic (16) is computed within the sub-DAG $\mathcal{G}^+_{\text{sub}}$, which reduces computation. Furthermore, Theorem 2 holds true when the estimator of the smaller precision matrix $\boldsymbol{\Omega}_{\text{sub}}$ enjoys the desired convergence rate

$O_{\mathbb{P}}(\sqrt{|S_{\text{sub}}|\log(p_{\text{sub}} \vee n)/n})$ in operator norm, where the subscript $_{\text{sub}}$ denotes the quantities corresponding to the structural equations of $\boldsymbol{Y}_{\text{sub}}$.

5 Theory

In this section, we develop a theory to quantify the finite sample performance as well as the complexities of Algorithms 1–2 when TLP is used for computation.

To proceed, we introduce some technical conditions for casual discovery consistency. For $(k, j) \in \mathcal{E}^+$, let $\boldsymbol{\Sigma}^{(k,j)}$ be the covariance matrix of $(\mathbb{E}(Y_k \mid \boldsymbol{Y}_{\text{nm}\mathcal{G}(k,j)}, \boldsymbol{X}), \boldsymbol{Y}_{\text{nm}\mathcal{G}(k,j)}, \boldsymbol{X})$. Moreover, let $s = \max_{(k,j)\in\mathcal{E}^+}(\kappa + \nu_2, \nu_1) \vee \max_{1\le k\le p} \|\mathbf{V}_{+,k}\|_0$ be the maximum sparsity-level in the estimation procedure, where ν_1, ν_2, κ depends on (*k*, *j*) which is dropped for conciseness. Assume there exist constants $c_0, c_1, c_2, c_3 > 0$ such that

(C1) $\min_{(k,j)\in\mathcal{E}^+} \min_{B:|B|\le 2s} \min_{\mathbf{v}:|\mathbf{v}|_2=1, |\mathbf{v}_{B^c}|_1 \le 3|\mathbf{v}_B|_1 + c_0 s\sqrt{\log(p)/n}} \langle \mathbf{v}, \boldsymbol{\Sigma}^{(k,j)}\mathbf{v}\rangle \ge c_1$.

(C2) $\min_{\mathrm{V}_{kj}\neq 0} |\mathrm{V}_{kj}| \ge c_2\sqrt{\log(q \vee n)/n}$.

(C3) $\min_{\mathrm{U}_{kj}\neq 0} |\mathrm{U}_{kj}| \ge c_3\sqrt{\log(p \vee n)/n}$.

(C4) $\max_{1\le k\le p}\{|\mathrm{an}_{\mathcal{G}}(k)|, |\mathrm{in}_{\mathcal{G}}(k)|, \|\mathbf{U}_{+,k}\|_1\} = O(1)$, and $\max_{(k,j)\in\mathcal{E}^+}(\mathrm{Diag}(\boldsymbol{\Sigma}^{(k,j)})) = O(1)$.

Condition (C1) is a restricted eigenvalue condition, which is common in high-dimensional estimation (Bickel et al., 2009) and can be viewed as a stronger version of (A1) in Theorem 1. (C2) and (C3) impose restrictions on the minimal signal strengths of $\mathbf{V}$ and $\mathbf{U}$ so that the ARG $\mathcal{G}^+$ and DAG $\mathcal{G}$ can be consistently recovered, respectively. They are similar to the beta-min condition (Meinshausen and Bühlmann, 2006) and the degree of separation condition (Shen et al., 2012) in the variable selection literature.

Theorem 3. *Suppose Assumptions (A1)–(A3) in Theorem 1 are satisfied and assume* $\boldsymbol{X}$ *is sub-Gaussian with mean zero and parameter* ς^2.

(A) (Parameter estimation) Suppose (C1), (C2), (C4) are met with sufficiently large c_0, c_1, c_2. *Suppose the tuning parameters are suitably chosen such that*

(I) *In Algorithm 1,* $0.01c_2\sqrt{\log(q \vee n)/n} \leq \tau' \leq 0.4 \min_{V_{kj} \neq 0} |V_{kj}|, \kappa'_j = |\mathbf{V}_{+,j}|_0$ *for* $1 \leq j \leq p$.

(II) *In Algorithm 2,*

$0.5c_3\sqrt{\log(p \vee n)/n} \leq \tau$, $\nu_1 = \lceil \mathrm{TLP}_\tau((\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) \rceil$, $\nu_2 = \lceil \mathrm{TLP}_\tau(\boldsymbol{\gamma}) \rceil$, *and* $\kappa = \lceil \mathrm{TLP}_\tau(\boldsymbol{\beta}) \rceil$ *for any* $(k, j) \in \mathcal{E}^+$.

Then there exists constant $C_1 > 0$ *such that when n is sufficiently large*

$$|U_{kj} - U_{kj}| \leq C_1\sqrt{\log(p \vee n)/n},$$

almost surely under $\mathbb{P}_{(\mathbf{U}, \mathbf{W}, \boldsymbol{\Sigma})}$. *Moreover, Algorithms 1 and 2 respectively terminate in* $O(p \times \log(s) \times (q^3 + nq^2))$ *and* $O(|\mathcal{E}^+| \times \log(s) \times (q^3 + nq^2))$ *operations almost surely.*

(B) *(Graph recovery) Additionally, if (C3) is satisfied with* $c_3 > C_1 > \tau$, *then when n is sufficiently large we have* $\mathcal{G} = \mathcal{G}$ *almost surely.*

By Theorem 3, the proposed method achieves causal discovery consistency in terms of consistent parameter estimation and structure recovery. Moreover, Algorithms 1–2 enjoy low-order polynomial time complexity almost surely provided that the data are randomly sampled from (1).

6 Numerical examples

6.1 Simulations

This subsection investigates via simulations the operating characteristics of GrIVET, including the qualities of structure learning, parameter estimation, and statistical inference.

To generate an observation $(\boldsymbol{Y}, \boldsymbol{X})$, we first introduce hidden variables $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{I}_{r \times r})$ as unmeasured confounders. Then, we sample $\boldsymbol{X}$ from $N(\mathbf{0}, \mathbf{I}_{q \times q})$ for continuous interventions or from $\{-1, 1\}^q$ with equal probability for discrete interventions. Given $\boldsymbol{X}$ and $\boldsymbol{\eta}$, we generate $\boldsymbol{Y}$ according to

$$\boldsymbol{Y} = \mathbf{U}^\top \boldsymbol{Y} + \mathbf{W}^\top \boldsymbol{X} + \boldsymbol{\Phi}^\top \boldsymbol{\eta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim N\left(\mathbf{0}, \mathrm{Diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)\right). \quad (17)$$

We conduct simulations with the following settings.

- **Hub graph.** Let p = 101, q = 252, and r = 10. For $\mathbf{U}, (\mathrm{U}_{1,j})_{2\le j\le p}$ are independently sampled from $\{-1,1\}$ with equal probability, while the rest are set to 0. This generates a sparse graph with the dense neighborhood of the first node. Let $\mathbf{W}_{q\times p} = (\mathbf{I}_{p\times p}, \mathbf{I}_{p\times p}, \mathbf{F}^{\top})^{\top}$ where the entries $(\mathrm{F}_{j,2j}, \mathrm{F}_{j,2j+1})_{1\le j\le q-2p}$ are set to 1, while other entries of **F** are zero. Then X_j, X_{2j} are IVs of Y_j for $j = 1,\dots,p$ and $X_{2p+1},\dots,X_q$ are invalid IVs with two intervention targets. For the confounders, $\Phi_{1,1}$ and $(\Phi_{jk})_{10j-8\le k\le 10j+1}^{1\le j\le r}$ are sampled uniformly from $(-0.4,-0.6)\cup(0.4,0.6)$, while other entries of $\mathbf{\Phi}$ are zero. We generate $(\sigma_1,\dots,\sigma_p)$ uniformly from $(0.4,0.6)$.
- **Random graph.** Let p = 100, q = 250, and r = 10. For $\mathbf{U}$, the upper off-diagonals $(\mathrm{U}_{kj})_{k<j}$ are sampled independently from {0, 1} according to $\mathrm{Bernoulli}(1/10p)$ while other entries are zero. Set $\mathbf{W}_{q\times p} = (\mathbf{I}_{p\times p}, \mathbf{I}_{p\times p}, \mathbf{F}^{\top})^{\top}$ where $(\mathrm{F}_{j,2j-1}, \mathrm{F}_{j,2j})_{1\le j\le 1-2p}$ are set to 1, while other entries of **F** are zero. Then X_j, X_{2j} are IVs of Y_j for $j = 1,\dots,p$ and $X_{2p+1},\dots,X_q$ are invalid IVs with two intervention targets. For the confounders, $(\Phi_{jk})_{10j-9\le k\le 10j}^{1\le j\le r}$ are sampled uniformly from $(-0.4,-0.6)\cup(0.4,0.6)$, while other entries of $\mathbf{\Phi}$ are zero. We generate $(\sigma_1,\dots,\sigma_p)$ uniformly from $(0.4,0.6)$.

Structure learning.

After obtaining ancestral relations from Algorithm 1, we implement Algorithm 2 to confirm parental relations but with constraints also imposed on the parameter of interest. Four graph metrics are used for evaluation: the false discovery rate (FDR), the true positive rate (TPR), the Jaccard index (JI), and the structural Hamming distance (SHD). The results in Table 1 demonstrate the strong performance of GrIVET in structure learning. Note that a high TPR indicates GrIVET's capability to detect the true existing edges, while the FDR remains low, signifying the high specificity of GrIVET. In Supplementary Materials Section 3.3, we further compare GrIVET with RFCI (Colombo et al., 2012) and LRpS-GES (Frot et al., 2019) in terms of structural learning accuracy. GrIVET compares favorably against the competitors.

Parameter estimation.

We compare the proposed IV estimation method in Section 3.3 with the regression method without any adjustment for confounding (Li et al., 2023a). To evaluate the quality of estimation, we consider three metrics, the average maximum absolute deviation, the mean absolute deviation, and the mean square deviation between true coefficients and estimates over 1000 runs. As demonstrated in Table 2, GrIVET enhances parameter estimation by accounting for latent confounding. As anticipated, GrIVET's estimation improves with increasing sample size n, while the naive regression method (Li et al., 2023a) remains inconsistent. Furthermore, GrIVET's advantages become more pronounced when stronger confounding effects are present, as evidenced by additional simulations in the Supplementary Materials.

Inference.

We now evaluate the empirical performance of the proposed tests in terms of size and power. For the empirical size, we calculate the percentage of times H_0 is rejected out of 1000 simulations when H_0 is true. For the power, we consider three alternative hypotheses H_a, where all the edges in H_0 exist. The empirical power of a test is the percentage of times H_0 is rejected out of 1000 simulations when H_a is true. The adjacency matrix $\mathbf{U}$ is modified according to the null and alternative hypotheses.

- **Hub graph, fixed $\mathcal{H}$.** For the size, consider $\mathcal{H} = \{(2,7)\}$, $\mathcal{H} = \{(2,7),(7,12),(12,17)\}$, and $\mathcal{H} = \{(2,7),(7,12),(12,17),(17,22),(22,27)\}$. For the power, consider $\mathcal{H} = \{(1,2)\}$, $\mathcal{H} = \{(1,2),(1,12),(1,22)\}$, and $\mathcal{H} = \{(1,2),(1,12),(1,22),(1,32),(1,42)\}$.
- **Random graph, fixed $\mathcal{H}$.** We consider $\mathcal{H} = \{(1,6)\}$, $\mathcal{H} = \{(1,6),(6,11),(11,16)\}$, and $\mathcal{H} = \{(1,6),(6,11),(11,16),(16,21),(21,26)\}$ for both size and power.
- **Random graph, random $\mathcal{H}$.** We also consider testing 50 randomly selected edges individually. Here, a random graph is generated so that 20 of these selected edges are present in the true DAG (i.e., H_a is valid). As a result, for every selected edge, H_0 holds in roughly 600 repetitions and H_a holds in roughly 400 repetitions.

As shown in Table 3 for fixed $\mathcal{H}$, empirical sizes are close to the nominal $\alpha = 0.05$ under H_0, and the proposed test enjoys desirable power under H_a. Figure 2 presents similar results for testing random $\mathcal{H}$. The Supplementary Materials display that the sampling distribution of the test statistic is close to the derived asymptotic distribution in Theorem 2. Additional simulation details and results are also available in Supplementary Materials.

6.2 ADNI data analysis

In this subsection, GrIVET is applied to analyze the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (available at https://adni.loni.usc.edu). The goal is to infer gene pathways related to Alzheimer's Disease (AD) in order to elucidate the gene-gene interactions in AD/cognitive impairment patients and healthy individuals, respectively.

Dataset.

The dataset comprises gene expression levels adjusted for five covariates: gender, handedness, education level, age, and intracranial volume. For data analysis, we select genes with at least one SNP at a marginal significance level below 10^{-14}, resulting in p = 21 genes as primary variables. For these genes, we further extract their marginally most correlated two SNPs, yielding q = 42 SNPs as unspecified intervention variables for subsequent data analysis. All gene expression levels are normalized.

The dataset initially categorizes individuals into four groups: Alzheimer's Disease (AD), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Cognitive Normal (CN). For our analysis, we treat 247 CN individuals as controls and the remaining 462 individuals as cases (AD-MCI). We then use the gene expressions and the SNPs to infer gene pathways for the 462 AD-MCI and 247 CN control cases, respectively.

Hypotheses.

We focus on statistical inferences related to genes APP and CASP3 (Julia and Goate, 2017; Su et al., 2001). As in Figure 3, for each edge (*k*, *j*), we consider testing $H_0 : \mathrm{U}_{kj} = 0$ versus $H_a : \mathrm{U}_{kj} \neq 0$.

Results.

Figure 3 displays the p-values and significant results under the level $\alpha = 0.05$ after the Holm-Bonferroni adjustment for $2 \times 7 = 14$ tests. The tests exhibit strong evidence for the presence of $\{\mathrm{LRP1} \rightarrow \mathrm{CASP3}, \mathrm{APP} \rightarrow \mathrm{APOE}\}$ in the AD-MCI group, but no evidence in the CN group. Meanwhile, this result suggests the presence of connections $\{\mathrm{CAPN1} \rightarrow \mathrm{CASP3}, \mathrm{ATP5F1} \rightarrow \mathrm{CASP3}\}$ in the CN group but not so in the AD-MCI group. In both groups, we identify directed connection $\mathrm{APP} \rightarrow \mathrm{APBB1}$. Figure 4 shows the residual correlation matrices for both groups, suggesting the existence of unmeasured confounding. The Supplementary Materials include normal Q-Q plots of residuals, demonstrating that the normality assumption is approximately satisfied for both groups.

Some of our discoveries agree with the existing findings. Specifically, our result indicates the presence of connection APP → APOE for the AD-MCI group, but not for the CN group, which seems consistent with the knowledge that APP and APOE are functionally linked in brain cholesterol metabolism (Liu et al., 2017) and the contributions of APOE to the pathophysiology of AD (Bu, 2009). The connection LRP1 → CASP3 also differs in AD-MCI and CN groups, which may serve to support the conclusion that activated CASP3 may be a factor in functional decline and may have an important role in neuronal cell death and plaque formation in AD brain (Su et al., 2001) given the finding that both APOE and its receptor LRP1 are present in amyloid plaques (Poirier, 1996). Moreover, the connection CAPN1 → CDK5R1 discovered in both groups can be found in the AlzNet database (interaction ID 24614).

7 Discussion

This article proposes a novel instrumental variable procedure that integrates causal discovery and inference for a Gaussian directed acyclic graph with hidden confounders. One future research direction is to develop methodologies for analyzing discrete/mixed-type (primary variable) data. Additionally, the present work uses individual-level data from a single study for causal discovery and inference. In many real applications, due to privacy concerns and ownership restrictions, the data are only available in the form of summary statistics (e.g., GWAS summary data) or in other privatized forms. Extending GrIVET to leverage these data is an important topic. Furthermore, multisource/decentralized data are ubiquitous, raising new challenges in communication, privacy, and handling of corrupted data. It would be promising to employ modern machine learning techniques, such as federated learning (Xiong et al., 2021; Gao et al., 2021), to address these challenges and fully unleash the potential of large-scale causal discovery and inference.

Finally, we discuss two limitations of the present work.

- GrIVET necessitates the availability of valid IVs for each primary variable due to the hardness of causal identification in the presence of hidden confounding. In genetic research, there is an ample supply of genetic variants (e.g., SNPs) serving as IVs. Nonetheless, obtaining valid IVs can be challenging in certain applications. It is thus crucial to investigate the potential for causal discovery even when faced with an insufficient number of IVs.
- For inference, Theorem 2 requires that $\mathbb{P}(\mathcal{G}^+ = \mathcal{G}^+) \to 1$, which is guaranteed by Condition (C2) in Theorem 3. Fulfilling this requirement can be challenging; in such cases, one might turn to the post-selection inference framework (Berk et al., 2013) by concentrating on the parameters within the selected model. However, the test results should be meticulously interpreted, as these parameters cease to be causal or structural (Berk et al., 2013) unless $\mathbb{P}(\mathcal{G}^+ = \mathcal{G}^+) \to 1$. In essence, (C2) enables the causal meaning of the tested parameters to be carried over to finite-sample inference. Exploring ways to lift the signal strength condition while preserving the causal interpretation for

statistical inference after DAG structure learning (Wang et al., 2023) is an important research topic.

A Appendix

Definition of d-separation (Pearl, 2009).

Consider a DAG $\mathcal{G}$ with node variables $(Z_1,\dots,Z_d)^\top$. Nodes Z_k and Z_j are adjacent if $Z_k \to Z_j$ or $Z_k \leftarrow Z_j$. An undirected path between Z_k and Z_j in $\mathcal{G}$ is a sequence of distinct nodes $(Z_k,\dots,Z_j)$ such that all pairs of successive nodes in the sequence are adjacent. A non-endpoint node Z_m on an undirected path $(Z_k,\dots,Z_{m-1},Z_m,Z_{m+1},\dots,Z_j)$ is called a collider if $Z_{m-1} \to Z_m \leftarrow Z_{m+1}$. Otherwise, it is called a non-collider. Let $A \subseteq \{1,\dots,d\}$, where A does not contain k and j. Then $\boldsymbol{z}_A$ is said to block an undirected path $(Z_k,\dots,Z_j)$ if at least one of the following holds: (1) the undirected path contains a non-collider that is in $\boldsymbol{z}_A$, or (2) the undirected path contains a collider that is not in $\boldsymbol{z}_A$ and has no descendant in $\boldsymbol{z}_A$. A node Z_k is d-separated from Z_j given $\boldsymbol{z}_A$ if $\boldsymbol{z}_A$ block every undirected path between Z_k and Z_j, $k \neq j$.

Additional discussion of Figure 1 (a).

Let $(k,j) \in \mathcal{E}^+$ and suppose all IVs are valid. We explain why $\boldsymbol{X}_{\mathrm{ca}(k)}$ may not be valid IVs after conditioning on $\boldsymbol{Y}_{\mathrm{an}(j)\setminus\{k\}}$, as mentioned in Section 3.3. Let $l \in \mathrm{ca}(k)$ and $m \in \mathrm{me}(k,j)$ such that Y_k is an unmediated parent of Y_m. Note that in Figure 1 (a) of the main text, whenever $\eta \to Y_m$, then $\boldsymbol{Y}_{\mathrm{an}(j)\setminus\{k\}}$ does not d-separate $\boldsymbol{X}_{\mathrm{ca}(k)}$ and η, since Y_m is a collider in the undirected path (X_l,Y_k,Y_m,η,Y_j). As a result, $\boldsymbol{X}_{\mathrm{ca}(k)}$ and η can be associated conditioned on $\boldsymbol{Y}_{\mathrm{an}(j)\setminus\{k\}}$.

Additional discussion on identification of $\mathbf{U}$.

We have the following result.

Lemma 1. *In (1), assume $\boldsymbol{X}$ and ε are independent.*

(A) $\mathbb{E}(Y_k \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X})$ *is a linear combination of* $(\boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X})$.

(B) $\mathbb{E}(\varepsilon_j \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X})$ *is a linear combination of* $(\boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}_{\mathrm{ca}(k)^c})$.

Proof. Here, (A) follows directly from (1). For (B), we have

$$\mathbb{E}(\varepsilon_j \mid \boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}) = \mathbb{E}(\varepsilon_j \mid \boldsymbol{\varepsilon}_{\mathrm{nm}(k,j)}, \boldsymbol{X}) = \mathbb{E}(\varepsilon_j \mid \boldsymbol{\varepsilon}_{\mathrm{nm}(k,j)}) = \boldsymbol{\pi}^\top \boldsymbol{\varepsilon}_{\mathrm{nm}(k,j)},$$

where the last equality is due to the normality of ε. Finally, in (1), we immediately have $\boldsymbol{\varepsilon}_{\mathrm{nm}(k,j)}$ is linear in $(\boldsymbol{Y}_{\mathrm{nm}(k,j)}, \boldsymbol{X}_{\mathrm{ca}(k)^c})$. □

Now, we show that $\mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = \mathbf{0}$ is sufficient to derive the identification results in Section 3.3. Given random variables ζ and ξ, let $\mathbb{L}(\zeta \mid \xi)$ be the best linear approximation of ζ using ξ, namely $\mathbb{L}(\zeta \mid \xi) = \boldsymbol{\omega}^\top \xi$ where

$$\boldsymbol{\omega} = \arg\min_{\omega} \mathbb{E}(\zeta - \boldsymbol{\omega}^\top \boldsymbol{\xi})^2.$$

For random variables ζ, ζ', and ξ, we have that (a) $\mathbb{L}(\zeta + \zeta' \mid \xi) = \mathbb{L}(\zeta \mid \boldsymbol{\xi}) + \mathbb{L}(\zeta' \mid \xi)$, (b) $\mathbb{L}(c\zeta \mid \boldsymbol{\xi}) = c\mathbb{L}(\zeta \mid \xi)$ for $c \in \mathbb{R}$, (c) $\mathbb{L}(\zeta \mid \boldsymbol{\xi}) = 0$ if $\mathrm{Cov}(\zeta, \xi) = \mathbf{0}$, (d) $\mathbb{L}(\zeta \mid \xi) = \zeta$ if $\zeta \in \mathrm{Span}(\xi)$, and (e) $\mathbb{L}(\zeta \mid \xi) = \mathbb{L}(\zeta \mid \mathbf{A}\xi)$ for invertible $\mathbf{A}$. Thus, $\mathbb{L}(\cdot \mid \star)$ mimics $\mathbb{E}(\cdot \mid \star)$, and Lemma 2 holds. The proof is similar to that of Lemma 1.

Lemma 2. *In (1), Lemma 1 holds with $\mathbb{E}(\cdot \mid \star)$ being replaced by $\mathbb{L}(\cdot \mid \star)$.*

As a result, if $\boldsymbol{X}$ and ε are uncorrelated as in (1), the derivation in Section 3.3 holds with $\mathbb{E}(\cdot \mid \star)$ being replaced by $\mathbb{L}(\cdot \mid \star)$.

Supplementary materials

Supplementary Materials include implementation details, additional simulations, and technical proofs.

ENDNOTES

[1]The causal parameter $\mathbf{U}$ is said to be identifiable if for any $(\mathbf{U}, \mathbf{W}, \boldsymbol{\Sigma})$ and $(\mathbf{U}', \mathbf{W}', \boldsymbol{\Sigma}')$, we have $\mathbb{P}_{\mathbf{U},\mathbf{W},\boldsymbol{\Sigma}} = \mathbb{P}_{\mathbf{U}',\mathbf{W}',\boldsymbol{\Sigma}'}$ implies $\mathbf{U} = \mathbf{U}'$. Otherwise, it is said to be non-identifiable.

[2]In Algorithm 1 Step 7, the indices of $\mathbf{V}$ are kept so that v_{lj} always represents the effect from X_l to Y_j.

References

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

Aragam, B., Amini, A. A., and Zhou, Q. (2019). Globally optimal score-based learning of directed acyclic graphs in high-dimensions. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 4450–4462.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

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

Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11(1):1–11.

Chakrabortty, A., Nandy, P., and Li, H. (2018). Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*.

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

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393.

Frot, B., Nandy, P., and Maathuis, M. H. (2019). Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):459–487.

Gao, E., Chen, J., Shen, L., Liu, T., Gong, M., and Bondell, H. (2021). FedDAG: Federated DAG structure learning. *Transactions on Machine Learning Research*.

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

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

Grimmer, J., Knox, D., and Stewart, B. M. (2020). Naïve regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *arXiv preprint arXiv:2007.12702*.

Gu, J., Fu, F., and Zhou, Q. (2019). Penalized estimation of directed acyclic graphs from discrete data. *Statistics and Computing*, 29(1):161–176.

Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815.

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

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

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

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144.

Kertel, M., Harmeling, S., and Pauly, M. (2022). Learning causal graphs in manufacturing domains using structural equation models. *arXiv preprint arXiv:2210.14573*.

Lee, K.-Y. and Li, L. (2022). Functional structural equation model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):600–629.

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

Li, C., Shen, X., and Pan, W. (2023a). Inference for a large directed acyclic graph with unspecified interventions. *Journal of Machine Learning Research*, 24(73):1–48.

Li, C., Shen, X., and Pan, W. (2023b). Nonlinear causal discovery with confounders. *Journal of the American Statistical Association*, pages 1–32.

Li, L., Shi, C., Guo, T., and Jagust, W. J. (2022). Sequential pathway inference for multimodal neuroimaging analysis. *Stat*, 11(1):e433.

Li, Y., Torralba, A., Anandkumar, A., Fox, D., and Garg, A. (2020b). Causal discovery in physical systems from videos. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 9180–9192.

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

Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, 15(1):1–7.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

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

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

Pearl, J. (2009). *Causality*. Cambridge University Press.

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

Poirier, J. (1996). Apolipoprotein E in the brain and its role in Alzheimer's disease. *Journal of Psychiatry and Neuroscience*, 21(2):128–134.

Rajendran, G., Kivva, B., Gao, M., and Aragam, B. (2021). Structure learning in polynomial time: Greedy algorithms, Bregman information, and exponential families. In *Advances in Neural Information Processing Systems*, volume 34, pages 18660–18672.

Reisach, A., Seiler, C., and Weichwald, S. (2021). Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Shah, R. D., Frot, B., Thanei, G.-A., and Meinshausen, N. (2020). Right singular vector projection graphs: fast high dimensional covariance matrix estimation under latent confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):361–389.

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

Shi, C. and Li, L. (2021). Testing mediation effects using logic of Boolean matrices. *Journal of the American Statistical Association*, pages 1–14.

Shi, C., Zhou, Y., and Li, L. (2023). Testing directed acyclic graph via structural, supervised and generative adversarial learning. *Journal of the American Statistical Association*, pages 1–24.

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

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

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

Vowels, M. J., Camgoz, N. C., and Bowden, R. (2021). D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys (CSUR)*.

Wang, Y. S., Kolar, M., and Drton, M. (2023). Confidence sets for causal orderings. *arXiv preprint arXiv:2305.14506*.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350.

Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., and Athey, S. (2021). Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*.

Xue, H. and Pan, W. (2020). Inferring causal direction between two traits in the presence of horizontal pleiotropy with GWAS summary data. *PLoS Genetics*, 16(11):e1009105.

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

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

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

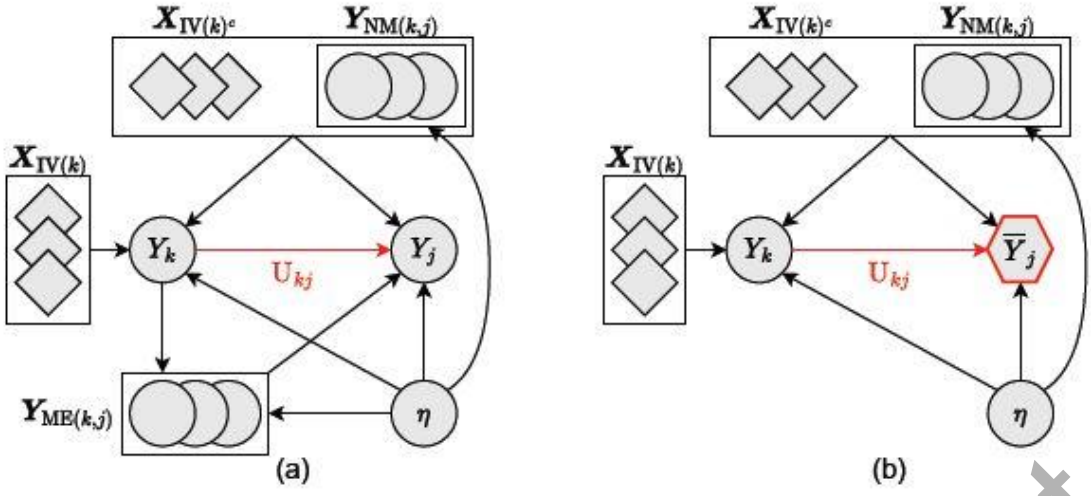


Fig. 1 Estimation of causal parameter U_{kj}. (a) Display of the relations among relevant variables. (b) Display of working response regression.

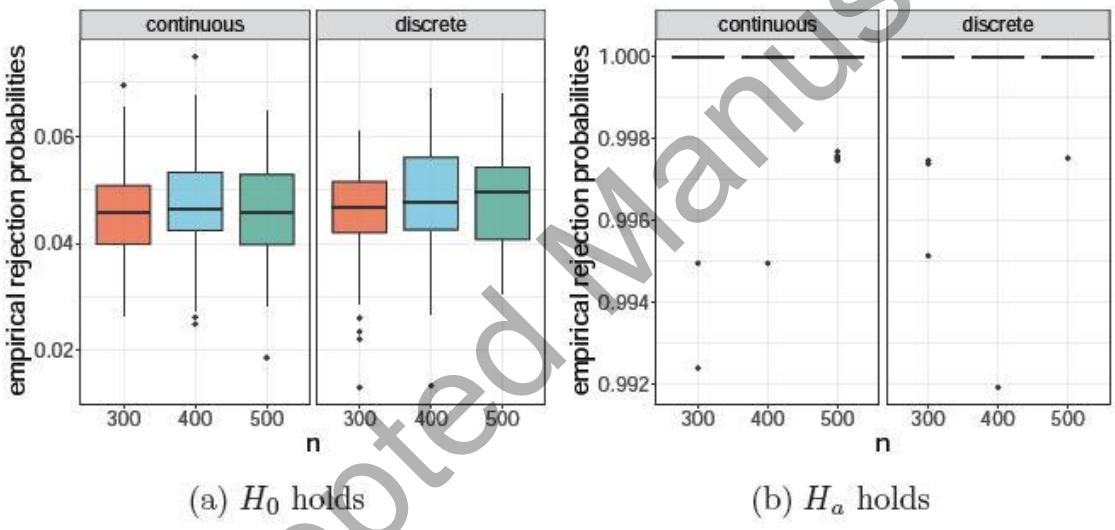
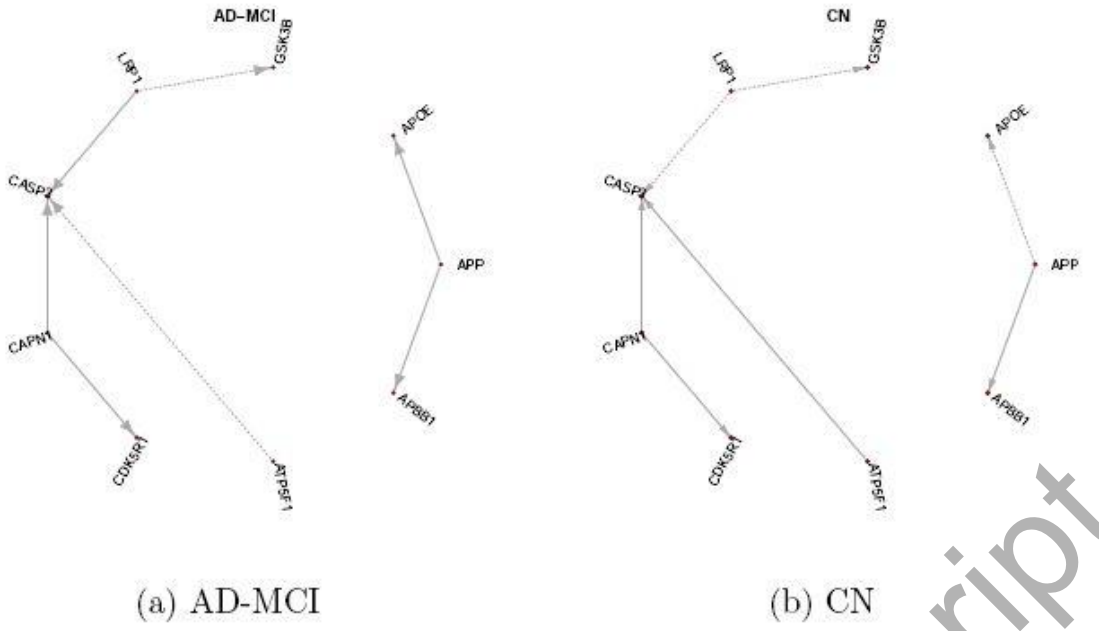
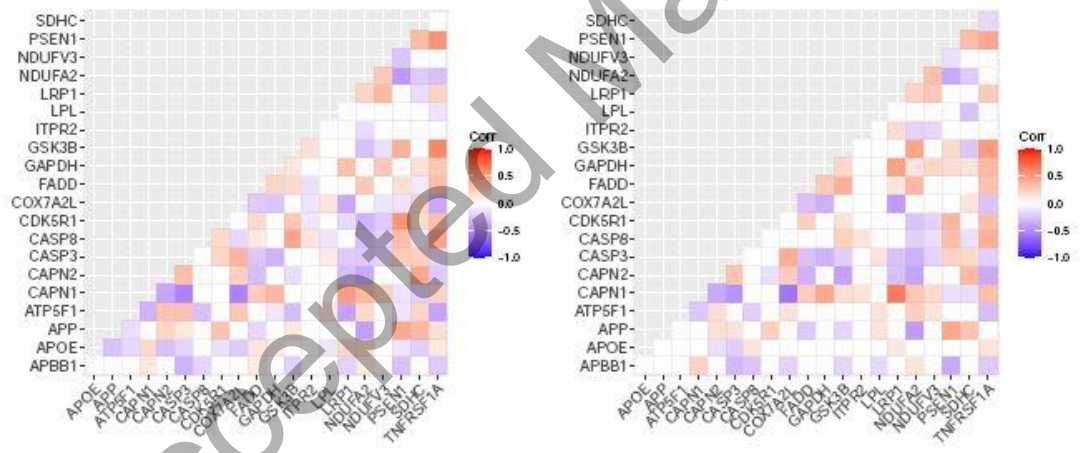


Fig. 2 The boxplots of the empirical rejection probabilities for testing randomly selected edges. The nominal level is $\alpha = 0.05$.



(a) AD-MCI (b) CN

Fig. 3 Display of the genes associated with proposed tests. (a) and (b): Solid/dashed arrows indicate significant/insignificant edges at $\alpha = 0.05$ after adjustment for multiplicity by the Bonferroni-Holm correction.



(a) AD-MCI (b) CN

Fig. 4 Display of residual correlation matrices for AD-MCI and CN groups.

Table 1 False discovery rate (FDR), true positive rate (TPR), structural Hamming distance (SHD), and Jaccard index (JI) of GrIVET for causal discovery over 1000 simulation replications. To compute the metrics, let TP, RE, FP, and FN be the numbers of identified edges with correct directions, those with wrong directions, estimated edges not in the skeleton of the true graph, and missing edges compared to the true skeleton. Then $FDR = (RE + FP)/(TP + RE + FP)$, $TPR = TP/(TP + FN)$, $SHD = FP + FN + RE$, and $JI = TP/(TP + SHD)$.

Graph	Intervention	n	FDR(%)	TPR(%)	SHD	JI(%)
Hub	Continuous	500	0.000	100.000	0.000	100.000
		400	0.000	99.998	0.002	99.998
		300	0.000	99.998	0.002	99.998
	Discrete	500	0.000	99.999	0.001	99.999
		400	0.000	99.998	0.002	99.998
		300	0.000	99.999	0.001	99.999
Random	Continuous	500	0.011	98.600	0.001	98.589
		400	0.000	98.600	0.000	98.600
		300	0.018	98.590	0.003	98.575
	Discrete	500	0.000	98.600	0.000	98.600
		400	0.024	98.600	0.002	98.576
		300	0.000	98.600	0.000	98.600

Table 2 Parameter estimation: the average of largest absolute difference (Max AD), the average absolute differences (Mean AD), and the average squared differences (Mean SqD) between the estimated parameters and the true parameters for two competing methods over 1000 simulation replications.

Graph	Intervention	*n*	GrIVET	Direct regression (Li et al., 2023a)
			(Max AD, Mean AD, Mean SqD)	(Max AD, Mean AD, Mean SqD)
Hub	Continuous	500	(0.06107, 0.01808, 0.00052)	(0.12817, 0.02448, 0.00142)
		400	(0.06863, 0.02037, 0.00066)	(0.13196, 0.02637, 0.00156)
		300	(0.07922, 0.02347, 0.00087)	(0.13395, 0.02873, 0.00170)
	Discrete	500	(0.06119, 0.01803, 0.00051)	(0.12770, 0.02434, 0.00141)
		400	(0.06932, 0.02030, 0.00065)	(0.13041, 0.02621, 0.00153)
		300	(0.08046, 0.02355, 0.00088)	(0.13334, 0.02867, 0.00169)
Random	Continuous	500	(0.02836, 0.01445, 0.00034)	(0.04254, 0.01791, 0.00076)
		400	(0.03245, 0.01660, 0.00045)	(0.04390, 0.01899, 0.00079)
		300	(0.03760, 0.01939, 0.00060)	(0.04709, 0.02150, 0.00091)
	Discrete	500	(0.02910, 0.01505, 0.00037)	(0.04287, 0.01808, 0.00075)
		400	(0.03272, 0.01686, 0.00046)	(0.04432, 0.01962, 0.00081)

Graph	Intervention	*n*	GrIVET	Direct regression (Li et al., 2023a)
		300	(0.03619, 0.01879, 0.00057)	(0.04756, 0.02146, 0.00094)

Table 3 Empirical size for GrIVET at nominal level $\alpha = 0.05$, respectively for $|\mathcal{D}| = 1, |\mathcal{D}| = 3$ and $|\mathcal{D}| = 5$, over 1000 simulation replications.

Graph	Intervention	*n*	Size ($\|\mathcal{D}\| = 1,3,5$)	Power ($\|\mathcal{D}\| = 1,3,5$)
Hub	Continuous	500	(0.028,0.026,0.029)	(1.000,1.000,1.000)
		400	(0.043,0.038,0.035)	(1.000,1.000,1.000)
		300	(0.037,0.030,0.034)	(1.000,1.000,1.000)
	Discrete	500	(0.036,0.040,0.027)	(1.000,1.000,1.000)
		400	(0.051,0.040,0.040)	(1.000,1.000,1.000)
		300	(0.052,0.041,0.035)	(1.000,1.000,1.000)
Random	Continuous	500	(0.038,0.037,0.026)	(1.000,1.000,1.000)
		400	(0.033,0.031,0.028)	(1.000,1.000,1.000)
		300	(0.033,0.025,0.030)	(1.000,1.000,1.000)
	Discrete	500	(0.040,0.029,0.027)	(1.000,1.000,1.000)
		400	(0.042,0.034,0.040)	(1.000,1.000,1.000)
		300	(0.029,0.033,0.034)	(1.000,1.000,1.000)