Semiparametric causal discovery and inference with invalid instruments

Jing Zou[1], Wei Li[2] and Wei Lin[1]

[1]*Peking University and* [2]*Renmin University of China*

Supplementary Material

This supplementary material includes examples, technical proofs and additional simulation results for the main text.

In this supplemental file, Section S1 includes some examples; Section S2 specifies the calculation of empirical distance correlation; Section S3 provides technical derivations and proofs of all theoretical results; Section S4 outlines the implementation details and extended results of the simulation studies in the main text; Section S5 provides additional simulations; Section S6 presents some additional analysis for the ADNI study.

S1 Examples

In this section, we present some examples for further illustration. We first consider Example S1.1 to illustrate several key concepts related to DAGs.

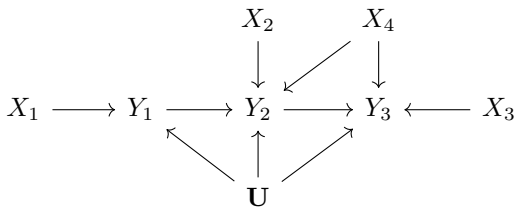


Figure S1.1: An example of the causal graph G.

Example S1.1. Consider the causal graph G shown in Figure S1.1, where **U** represents unobserved confounders. In view of the directed path $Y_1 \to Y_2 \to Y_3$, the mediator set of Y_1 and Y_3 is $\mathrm{me}_G(1,3) = \{2\}$. The unmediated parent of Y_3 is Y_2 because Y_2 is the only parent of Y_3 with $\mathrm{me}_G(2,3) = \emptyset$. The height of Y_1 is 2 since the longest path from Y_1 to a leaf node of G is $Y_1 \to Y_2 \to Y_3$, whose length is 2. The corresponding ancestral relation graph (ARG) is $G^+ = (\{X_1, \ldots, X_4\}, \{Y_1, Y_2, Y_3\}; \mathcal{E}^+, \mathcal{I}^+)$, where $\mathcal{E}^+ = \{(1,2),(1,3),(2,3)\}$ and $\mathcal{I}^+ = \{(1,1),(1,2),(1,3),(2,2),(2,3),(3,3),(4,2),(4,3)\}$. As shown in this example, the ARG describes the ancestral relationships among the nodes in G.

Since the estimation of $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$ plays an important role in our estimates for $\boldsymbol{\beta}^*$, we further clarify its concrete instantiation through Example S1.2.

Example S1.2. Focus on $k = 2$ in Figure S1.1, where $\gamma = 1$ according to Assumption 3. The candidate IV set for Y_2 is $\{2, 4\}$ and suppose that

$X_i \in \{0,1\},\ i = 2,4$. Then we have

$$\mathcal{H}(\mathbf{X}_{\{2,4\}}) = \text{span}(\{X_2 - \mu_2, X_4 - \mu_4, (X_2 - \mu_2)(X_4 - \mu_4)\}),$$

where $\mu_i = E(X_i),\ i = 2,4$. The valid IV set for Y_2 which satisfies $\text{iv}_G(2) \subseteq \text{ca}_G(2)$ and $|\,\text{iv}_G(2)| \geq 1$ can be $\{2\}$, $\{4\}$ or $\{2,4\}$, corresponding to the potential values for α_2 in Definition 3. According to Assumption 1, when $\alpha_2 = \{2\}$, the corresponding $\mathcal{D}(\{2\})$ is $\text{span}(\{X_2 - \mu_2, (X_2 - \mu_2)(X_4 - \mu_4)\})$; when $\alpha_2 = \{4\}$, $\mathcal{D}(\{4\}) = \text{span}(\{X_4 - \mu_4, (X_2 - \mu_2)(X_4 - \mu_4)\})$; and when $\alpha_2 = \{2,4\}$, $\mathcal{D}(\{2,4\}) = \text{span}(\{X_2 - \mu_2, X_4 - \mu_4, (X_2 - \mu_2)(X_4 - \mu_4)\})$. The intersection of all possible $\mathcal{D}(\alpha_2)$, defined as $\mathcal{Z}_1(2)$, is $\text{span}(\{(X_2 - \mu_2)(X_4 - \mu_4)\})$. Therefore, the vector $\mathbf{Z}_1(\mathbf{X}_{\text{ca}_G(2)})$ consisting of the basis functions of $\mathcal{Z}_1(2)$ is $(X_2 - \mu_2)(X_4 - \mu_4)$.

To further clarify the role of Assumption 1, we next provide an example where the identification fails specifically because this assumption is violated. In particular, we show that if Assumption 1 is violated, then the surrogate IVs may become invalid. As a result, the causal effects cannot be identified by (3.2), thereby invalidating the identification result in Theorem 1.

Example S1.3. Consider the causal graph G shown in Figure S1.2, where $\mathbf{U}$ represents unobserved confounders, and the dashed undirected edge between X_1 and X_2 indicates that $X_1 \not\perp\!\!\!\perp X_2$. In this graph, $|\,\text{ca}_G(j)| =$

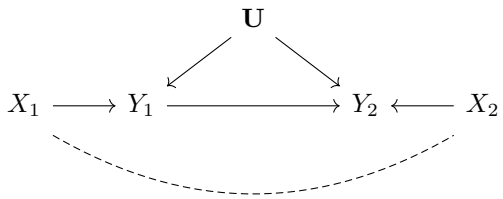


Figure S1.2: An example of the causal graph G when Assumption 1 fails.

$|\operatorname{iv}_G(j)| = \gamma = 1$ for $j = 1, 2$, and the corresponding SEM in (2.2) is

$$Y_1 = g_1(X_1) + \varepsilon_1,$$

$$Y_2 = \beta_{12}^* Y_1 + g_2(X_2) + \varepsilon_2.$$

We proceed to show that even in this simple case where $\operatorname{me}_G(1, 2) = \emptyset$, the surrogate IV denoted by $\mathbf{Z}_1(X_1)$ can be invalid, thereby causing (3.2) to fail. Suppose that $X_1 \in \{0, 1\}$ for simplicity. By Definition 3, $\mathbf{Z}_1(X_1) = X_1 - E(X_1)$, and thus

$$\begin{aligned} E\{\mathbf{M}(\boldsymbol{\beta}^*)\} &= E\{\mathbf{Z}_1(X_1)(Y_2 - \beta_{12}^* Y_1)\} \\ &= E\{(X_1 - E(X_1))(g_2(X_2) + \varepsilon_2)\} \\ &= E\{(X_1 - E(X_1))g_2(X_2)\}, \end{aligned}$$

which may not be zero due to the potential dependency between X_1 and X_2 and the unknown form of $g_2(\cdot)$. For example, if X_1 is correlated with X_2 and $g_2(X_2)$ is linear in X_2, then $E\{\mathbf{M}(\boldsymbol{\beta}^*)\} \neq \mathbf{0}$. As a result, the direct causal effect of Y_1 on Y_2 cannot be identified by (3.2).

S2 Distance correlation

In this section, we introduce the distance correlation (DC) proposed by Székely et al. (2007) in a brief. Let $f_\zeta(\cdot)$ denote the characteristic function for a random variable ζ.

Definition S2.1 (Distance correlation)**.** The distance correlation between random vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ with finite first moments is the non-negative number $\mathcal{R}(\boldsymbol{\xi}, \boldsymbol{\eta})$ defined by

$$\mathcal{R}^2(\boldsymbol{\xi}, \boldsymbol{\eta}) = \begin{cases} \frac{\mathcal{V}^2(\boldsymbol{\xi}, \boldsymbol{\eta})}{\sqrt{\mathcal{V}^2(\boldsymbol{\xi}, \boldsymbol{\xi})\mathcal{V}^2(\boldsymbol{\eta}, \boldsymbol{\eta})}} & , \ \mathcal{V}^2(\boldsymbol{\xi}, \boldsymbol{\xi})\mathcal{V}^2(\boldsymbol{\eta}, \boldsymbol{\eta}) > 0 \\ 0 & , \ \mathcal{V}^2(\boldsymbol{\xi}, \boldsymbol{\xi})\mathcal{V}^2(\boldsymbol{\eta}, \boldsymbol{\eta}) = 0 \end{cases}, \quad \text{where}$$

$$\mathcal{V}^2(\boldsymbol{\xi}, \boldsymbol{\eta}) = \|f_{\boldsymbol{\xi},\boldsymbol{\eta}}(t, s) - f_{\boldsymbol{\xi}}(t) f_{\boldsymbol{\eta}}(s)\|^2 = \frac{1}{c_1 c_2} \int_{\mathbb{R}^{p_{\boldsymbol{\xi}}+p_{\boldsymbol{\eta}}}} \frac{|f_{\boldsymbol{\xi},\boldsymbol{\eta}}(t, s) - f_{\boldsymbol{\xi}}(t) f_{\boldsymbol{\eta}}(s)|^2}{||t||_2^{1+p_{\boldsymbol{\xi}}} ||s||_2^{1+p_{\boldsymbol{\eta}}}} dt ds,$$

$\mathcal{V}^2(\boldsymbol{\xi}, \boldsymbol{\xi})$ and $\mathcal{V}^2(\boldsymbol{\eta}, \boldsymbol{\eta})$ are defined similarly, the non-negative number $\mathcal{V}(\boldsymbol{\xi}, \boldsymbol{\eta})$ is called distance covariance; the positive constants c_1 and c_2 only depend on the dimensions $p_{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$ and $p_{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$, respectively.

The distance correlation can be intuitively thought of as measuring the difference between the characteristic functions of the distribution under the assumption of independence of two random vectors, and the true joint one. Therefore, unlike the Pearson correlation coefficient, the distance correlation is not limited to measuring only linear relationships.

We next introduce the calculation of empirical DC in our setting. For finite sample estimation, consider an independent and identically distributed sample $(\mathbf{X}_{n\times q}, \mathbf{Y}_{n\times p})$. The empirical DC between X_i and Y_j is defined by

$$\mathcal{R}_n(X_i, Y_j) = \frac{\mathcal{V}_n(X_i, Y_j)}{\sqrt{\mathcal{V}_n(X_i, X_i)\mathcal{V}_n(Y_j, Y_j)}}, \tag{S2.1}$$

where $\mathcal{V}_n(\cdot,\cdot)$ is the empirical distance covariance. Note from Székely et al. (2007, (2.18)) that $\mathcal{V}_n^2(X_i, Y_j) = S_1(X_i, Y_j) + S_2(X_i, Y_j) - 2S_3(X_i, Y_j)$, where

$$\begin{aligned}
S_1(X_i, Y_j) &= \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} |X_{ri} - X_{si}|\,|Y_{rj} - Y_{sj}|,\\
S_2(X_i, Y_j) &= \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} |X_{ri} - X_{si}| \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} |Y_{rj} - Y_{sj}|,\\
S_3(X_i, Y_j) &= \frac{1}{n^3}\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n} |X_{ri} - X_{ti}|\,|Y_{sj} - Y_{tj}|.
\end{aligned}$$

The distance variances $\mathcal{V}_n(X_i, X_i)$ and $\mathcal{V}_n(Y_j, Y_j)$ are calculated similarly. For testing the null hypothesis $H_0\colon X_i \perp\!\!\!\perp Y_j$, the test statistic is given by

$$T_n(X_i, Y_j) = \frac{n\mathcal{V}_n^2(X_i, Y_j)}{S_2(X_i, Y_j)}.$$

Then a test of asymptotic significance level α rejects H_0, denoting $R_{ij} = 1$, where

$$R_{ij} := \mathbf{1}\left\{\sqrt{T_n(X_i, Y_j)} > \Phi^{-1}(1 - \alpha/2)\right\}, \tag{S2.2}$$

with $\Phi(\cdot)$ being the standard normal cumulative distribution function (Székely et al., 2007, Theorem 6).

Székely et al. (2007) also presented an alternative definition of empirical distance covariance as follows, which is equivalent to the aforementioned definition but more convenient for calculations. Considering the empirical DC between X_ℓ and Y_k, define

$$a_{ij} = |X_{i\ell} - X_{j\ell}|, \quad \bar{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \bar{a}_{\cdot_j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij},$$

$$\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}, \quad A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot},$$

$i, j = 1, \ldots, n$, $\mathbf{A} = (A_{ij})_{n\times n}$. Similarly, define $b_{ij} = |Y_{ik} - Y_{jk}|_2$, $B_{ij} = b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}, i, j = 1, \ldots, n$, $\mathbf{B} = (B_{ij})_{n\times n}$. The empirical distance covariance $\mathcal{V}_n(X_\ell, Y_k)$ is then defined as the non-negative number by

$$\mathcal{V}_n^2(X_\ell, Y_k) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}.$$

We next recall the asymptotic properties of the empirical DC and DC-based independence tests from Székely et al. (2007), which will be used to prove the consistency of the proposed ARG estimator in Algorithm 1.

Lemma S2.1 (Székely et al. (2007, Corollary 1 and Theorem 3))**.** *If the random variables $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ satisfy $E(\|\boldsymbol{\xi}\|_2 + \|\boldsymbol{\eta}\|_2) < \infty$, then almost surely,*

$$\lim_{n\to\infty} \mathcal{R}_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = \mathcal{R}(\boldsymbol{\xi}, \boldsymbol{\eta}),$$

where $0 \le \mathcal{R}(\boldsymbol{\xi}, \boldsymbol{\eta}) \le 1$, and $\mathcal{R}(\boldsymbol{\xi}, \boldsymbol{\eta}) = 0$ if and only if $\boldsymbol{\xi} \perp\!\!\!\perp \boldsymbol{\eta}$.

For testing $H_0 : \boldsymbol{\xi} \perp\!\!\!\perp \boldsymbol{\eta}$ versus $H_1 : \boldsymbol{\xi} \not\perp\!\!\!\perp \boldsymbol{\eta}$, let $P^I(n)$ denote the

Type I error probability of the DC-based test that rejects independence if $\sqrt{T_n(\boldsymbol{\xi}, \boldsymbol{\eta})} \geq \Phi^{-1}(1 - \alpha/2)$ for some significance level α.

Lemma S2.2 (Székely et al. (2007, Theorem 6))**.** *If the random variables* $\boldsymbol{\xi}$ *and* $\boldsymbol{\eta}$ *satisfy* $E(\|\boldsymbol{\xi}\|_2 + \|\boldsymbol{\eta}\|_2) < \infty$*, then for all* $0 < \alpha \leq 0.215$*,* $\lim_{n\to\infty} P^I(n) \leq \alpha$*.*

Building on these results, we show that Algorithm 1 with the significance level set to $\alpha = O(n^{-2})$ asymptotically yields the correct decision for every test $H_0 : X_i \perp\!\!\!\perp Y_j$ versus $H_1 : X_i \not\perp\!\!\!\perp Y_j$, where $i = 1, \ldots, q$ and $j = 1, \ldots, p$. Indeed, Lemma S2.1 implies that when $X_i \not\perp\!\!\!\perp Y_j$, the test statistic $T_n(X_i, Y_j)/n \to C$ for some $C > 0$ depending on $\mathcal{R}(X_i, Y_j)$. Therefore, the Type II error probability of the test vanishes asymptotically with $\alpha = O(n^{-2})$. On the other hand, Lemma S2.2 implies that if α tends to zero, then the Type I error probability also converges to zero. Consequently, with this choice of α, Algorithm 1 correctly recovers all dependencies between $\mathbf{X}$ and $\mathbf{Y}$ as $n \to \infty$, thereby consistently estimating the ARG.

S3 Proofs

In this section, we provide the proofs of Theorems 1–4. In particular, the proof strategy of Theorem 1 aligns with the estimation approaches intro-

duced in the main text.

S3.1 Proof of Theorem 1

Our identification strategy consists of two stages. We begin by identifying the ARG to roughly capture the causal directions among $\mathbf{Y}$ and obtain the candidate IV sets. We then establish the identification of the causal effects $\mathbf{B}^*$ based on the identified ARG and candidate IV sets.

First, we introduce some definitions. Let $l_G(k, j)$ denote the length of the longest directed path from Y_k to Y_j. Define the height of Y_j, $h_G(j)$, as the length of the longest directed path from Y_j to a leaf node of G. It follows that if $(k, j) \in \mathcal{E}$ then $h_G(k) > h_G(j)$, and the height of any leaf node is 0. For a DAG G, we define its ARG formally as follows.

Definition S3.1 (Ancestral relation graph)**.** For a causal graph $G = (\mathbf{X}, \mathbf{Y}; \mathcal{E}, \mathcal{I})$, its ancestral relation graph is defined by $G^+ = (\mathbf{X}, \mathbf{Y}; \mathcal{E}^+, \mathcal{I}^+)$, where

$$\mathcal{E}^+ = \{(k, j) : k \in \mathrm{an}_G(j)\}, \quad \mathcal{I}^+ = \left\{(\ell, j) : \ell \in \bigcup_{k \in \mathrm{an}_G(j) \cup \{j\}} \mathrm{in}_G(k)\right\}.$$

The ARG G^+ describes the ancestral relationships among the nodes in G. Specifically, if there exists a directed path from Y_i to Y_j in G, then $(i, j) \in \mathcal{E}^+$. Similarly, if there exits a directed path from X_ℓ to Y_j in G, then $(\ell, j) \in \mathcal{I}^+$. To recover $\mathcal{E}^+$, we need only identify all edges originating from the unmediated parents of each node in G. Note that we can derive

the mediator sets $\{\text{me}_G(k,j)\}_{(k,j)\in\mathcal{E}^+}$ and the lengths $\{l_G(k,j)\}_{(k,j)\in\mathcal{E}^+}$ from G^+ since $\text{me}_G(k,j) = \text{me}_{G^+}(k,j)$ and $l_G(k,j) = l_{G^+}(k,j)$.

The next two propositions extend Propositions 1 and 2 in Chen et al. (2024) to the semiparametric model (2.2) and are the key ingredients for identifying G^+ and $\{ca_G(k)\}_{k=1}^p$.

Proposition S3.1. *Suppose that Assumptions 1–3 hold. Then there exists some X_ℓ such that $X_\ell \not\perp\!\!\!\perp Y_k$ and $X_\ell \perp\!\!\!\perp Y_{k'}$ for all $k' \neq k$ if and only if Y_k is a leaf node of G. Furthermore, such an X_ℓ is a valid IV for Y_k in G.*

Proof of Proposition S3.1. On the one hand, if Y_k is a leaf node of G, then let X_ℓ be a valid IV of Y_k since $\text{iv}(k) \neq \emptyset$ due to Assumption 3. We then have $X_\ell \perp\!\!\!\perp Y_{k'}$ for any $k' \neq k$ because Y_k has no descendant and $X_\ell \perp\!\!\!\perp X_i$ for any $i \neq \ell$ according to Assumption 1. On the other hand, if Y_k is not a leaf node of G, then there exists a Y_j such that Y_k is an unmediated parent of Y_j. There are two cases leading to $X_\ell \not\perp\!\!\!\perp Y_k$, i.e., $(\ell,k) \in \mathcal{I}$ and $(\ell,k) \notin \mathcal{I}$. If $(\ell,k) \in \mathcal{I}$, then we have $X_\ell \not\perp\!\!\!\perp Y_j$ according to Assumption 2. If $(\ell,k) \notin \mathcal{I}$, then there exists an ancestor $i \in \text{an}_G(k)$ such that $(\ell,i) \in \mathcal{I}$, and thus $X_\ell \not\perp\!\!\!\perp Y_i$. Therefore, for any non-leaf of G, we can not find an X_ℓ satisfying the conditions in Proposition S3.1. The analysis above shows the identification of leaf nodes in G. Moreover, under Assumptions 1–3, for any leaf node Y_k of G, it follows that $X_\ell \not\perp\!\!\!\perp Y_k$ and $X_\ell \perp\!\!\!\perp Y_{k'}$ for all $k' \neq k$

if and only if $\ell \in \mathrm{iv}_G(k)$. □

Proposition S3.1 shows that the leaves of G and their valid IVs are identifiable:

$$\mathrm{leaf}(G) = \{k : \text{for some } \ell,\ X_\ell \not\perp\!\!\!\perp Y_k \text{ and } X_\ell \perp\!\!\!\perp Y_{k'} \text{ for all } k' \neq k\},$$

$$\mathrm{iv}_G(k) = \{\ell : X_\ell \not\perp\!\!\!\perp Y_k \text{ and } X_\ell \perp\!\!\!\perp Y_{k'} \text{ for all } k' \neq k\}, \quad k \in \mathrm{leaf}(G).$$

Once the leaves of G have been identified, we can remove these nodes along with their valid IVs to obtain a subgraph $G^- = (\mathbf{X}^-, \mathbf{Y}^-; \mathcal{E}^-, \mathcal{I}^-)$, where $\mathbf{X}^- = \mathbf{X} \setminus \bigcup_{k \in \mathrm{leaf}(G)} \mathbf{X}_{\mathrm{iv}_G(k)}$, $\mathbf{Y}^- = \mathbf{Y} \setminus \mathbf{Y}_{\mathrm{leaf}(G)}$, and $\mathcal{E}^-$ and $\mathcal{I}^-$ denote the remaining edges from $\mathcal{E}$ and $\mathcal{I}$, respectively. By Definition 1, it is clear that $\mathrm{iv}_G(j) \subseteq \mathrm{iv}_{G^-}(j)$ for all $Y_j \in \mathbf{Y}^-$, implying that Assumption 3 holds in G^-. Assumptions 1 and 2 also hold naturally in the subgraph G^-. Therefore, Proposition S3.1 remains applicable to G^-, so that $\mathrm{leaf}(G^-)$ and $\mathrm{iv}_{G^-}(k)$ for $k \in \mathrm{leaf}(G^-)$ are identifiable. By iteratively applying this method to identify and remove the leaves of the current graph, a topological order among the variables in $\mathbf{Y}$ can be determined. During this process, the variables in $\mathbf{Y}$ are removed in ascending order of their heights. It is obvious from the definition of height that there are no directed paths from Y_j to nodes with the same or greater height. However, the causal relationships for the other case are yet to be determined. The following proposition helps

to complete the construction of G^+.

Proposition S3.2. *Suppose that Assumptions 1–3 hold. For any* $k \in \text{leaf}(G^-)$ *and* $Y_j \in \mathbf{Y} \setminus \mathbf{Y}^-$*: (i) if* $X_\ell \not\perp\!\!\!\perp Y_j$ *for all* $\ell \in \text{iv}_{G^-}(k)$*, then* $(k, j) \in \mathcal{E}^+$*; (ii) if* Y_k *is an unmediated parent of* Y_j*, then* $X_\ell \not\perp\!\!\!\perp Y_j$ *for all* $\ell \in \text{iv}_{G^-}(k)$*.*

Proof of Proposition S3.2. It is obvious that $\text{iv}_G(k) \subseteq \text{iv}_{G^-}(k)$ for any $Y_k \in \mathbf{Y}^-$ according to Definition 1 for valid IVs. According to Assumption 1 and Definition 1, if $X_\ell \perp\!\!\!\perp Y_k$ for all $\ell \in \text{iv}_G(k)$, there must be a directed path from Y_k to Y_j in G. Thereby, we prove the first conclusion of Proposition S3.2. Moreover, for any $\ell \in \text{iv}_{G^-}(k)$, we have $(\ell, k) \in \mathcal{I}^- \subseteq \mathcal{I}$ based on the definition of valid IV in Definition 1. Therefore, if Y_k is an unmediated parent of Y_j, we have $X_\ell \not\perp\!\!\!\perp Y_j$ according to Assumption 2, which leads to the second conclusion. □

Proposition S3.2 allows us to derive the ancestral relations between $\mathbf{Y}_{\text{leaf}(G^-)}$ and $\mathbf{Y} \setminus \mathbf{Y}^-$ by

$$\{(k, j) : k \in \text{leaf}(G^-), Y_j \in \mathbf{Y} \setminus \mathbf{Y}^-, X_\ell \not\perp\!\!\!\perp Y_j \text{ for all } \ell \in \text{iv}_{G^-}(k)\} \subseteq \mathcal{E}^+,$$

which ensures that all edges from an unmediated parent to Y_j are included. The remaining edges in $\mathcal{E}^+$ correspond to the directed paths containing mediators in G. Since these paths are formed by edges previously identi-

fied, adding the ancestral relationships inferred from these paths to $\mathcal{E}^+$ is sufficient to recover $\mathcal{E}^+$. Moreover, we can reconstruct $\mathcal{I}^+$ by

$$\mathcal{I}^+ = \{(\ell, j) : \text{for some } k \in \text{an}_G(j) \cup \{j\}, X_\ell \not\perp\!\!\!\perp Y_k\}.$$

Subsequently, by Definition 2, the candidate IV sets are identified by

$$\text{ca}_G(k) = \{\ell : (\ell, k) \in \mathcal{I}^+ \text{ and } (\ell, j) \in \mathcal{I}^+, k \neq j \text{ only if } (k, j) \in \mathcal{E}^+\},\ k = 1, \ldots, p.$$

Building on the identifiability of $\mathcal{E}^+$ and $\{\text{ca}_G(k)\}_{k=1}^p$ established in the preceding subsection, we now proceed to demonstrate the identification of the causal effects $\boldsymbol{\beta}^* = (\beta^*_{kj})_{(k,j)\in\mathcal{E}^+}$ and subsequently identify $\mathcal{E}$.

First, consider the simple case where $(k, j) \in \mathcal{E}^+$ and $\text{me}_G(k, j) = \emptyset$. When Assumption 1 holds, we have $\mathbf{Y}_{\text{nm}_G(k,j)} \perp\!\!\!\perp \mathbf{X}_{\text{ca}_G(k)}$ and $\mathbf{X}_{\text{in}_G(j)} \perp\!\!\!\perp \mathbf{X}_{\text{ca}_G(k)} \mid \mathbf{X}_{\text{ca}_G(k)\setminus\text{iv}_G(k)}$. It then follows that for any $d(\mathbf{X}_{\text{ca}_G(k)}) \in \mathcal{D}(\text{iv}_G(k))$,

$$\begin{aligned}
&E\{d(\mathbf{X}_{\text{ca}_G(k)})(Y_j - \beta^*_{kj} Y_k)\} \\
&\quad = E\left\{ d(\mathbf{X}_{\text{ca}_G(k)}) \left(\sum_{i\in\text{nm}_G(k,j)} \beta^*_{ij} Y_i + g_j(\mathbf{X}_{\text{in}_G(j)}) + \varepsilon_j \right) \right\} \\
&\quad = E\{d(\mathbf{X}_{\text{ca}_G(k)}) g_j(\mathbf{X}_{\text{in}_G(j)})\} \\
&\quad = E\big[E\{d(\mathbf{X}_{\text{ca}_G(k)}) \mid \mathbf{X}_{\text{ca}_G(k)\setminus\text{iv}_G(k)}\} E\{g_j(\mathbf{X}_{\text{in}_G(j)}) \mid \mathbf{X}_{\text{ca}_G(k)\setminus\text{iv}_G(k)}\}\big] \\
&\quad = 0.
\end{aligned} \tag{S3.1}$$

By Definition 3 and the identifiability of $\text{ca}_G(k)$, $\mathcal{Z}_\gamma(k)$ and hence $\mathbf{Z}_\gamma(\mathbf{X}_{\text{ca}_G(k)})$ are also identifiable. Therefore, under Assumption 4, equation 3.1 has a

unique solution, and thus β^*_{kj} is identifiable. Based on the above analysis, we can identify all β^*_{kj} with $l_G(k,j) = 1$.

Next, we identify the remaining parameters recursively. Suppose we have identified all β^*_{kj} with $l_G(k,j) \leq l$ for some $l > 0$. Then, for any β^*_{kj} with $l_G(k,j) = l+1$, any mediator variable $Y_i \in \mathrm{me}_G(k,j)$ satisfies $l_G(i,j) \leq l$, and thus all mediated effects β^*_{ij} are identified. We can then substitute $Y_j - \sum_{i \in \mathrm{me}_G(k,j)} \beta^*_{ij} Y_i$ for Y_j and identify β^*_{kj} under Assumption 4 from the equation $E\{M_{kj}(\boldsymbol{\beta}^*)\} = 0$, where

$$M_{kj}(\boldsymbol{\beta}^*) = \mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)}) \left(Y_j - \sum_{i \in \mathrm{me}_G(k,j)} \beta^*_{ij} Y_i - \beta^*_{kj} Y_k \right).$$

The above procedure implies that we can determine all β^*_{kj} recursively in ascending order of $l_G(k,j)$. As a result, $\boldsymbol{\beta}^*$ can be identified as the unique solution to

$$E\{\mathbf{M}(\boldsymbol{\beta}^*)\} = \mathbf{0}, \tag{S3.2}$$

where $\mathbf{M}(\boldsymbol{\beta}^*)$ is the concatenation of all $M_{kj}(\boldsymbol{\beta}^*)$ for $(k,j) \in \mathcal{E}^+$. Finally, based on the identified value of $\boldsymbol{\beta}^*$, we can identify $\mathcal{E}$ by $\mathcal{E} = \{(k,j) : \beta^*_{kj} \neq 0, (k,j) \in \mathcal{E}^+\}$.

To summarize, our analysis shows that the causal graph and causal effects in model (2.2) are identifiable under Assumptions 1–4, thereby justifying Theorem 1.

S3.2 Proof of Theorem 2

Proof. Let $P_{ij}^{I}(n)$ and $P_{ij}^{II}(n)$ denote the probability of making a Type I and Type II error of the DC-based independence test of X_i and Y_j, respectively. Firstly, we prove that if we take $\alpha = O(n^{-2})$, then $\lim_{n\to+\infty} P_{ij}^{I}(n) = \lim_{n\to+\infty} P_{ij}^{II}(n) = 0$. Note that $E(X_i^2) < \infty$, $E(Y_j^2) < \infty$, for all $i = 1, \ldots, q, j = 1, \ldots, p$, and when n is sufficiently large, $\alpha \leq 0.215$. Therefore, by Theorem 6 of Székely et al. (2007), under the null hypothesis,

$$\lim_{n\to+\infty} P_{ij}^{I}(n) \leq \lim_{n\to+\infty} \alpha = 0.$$

If the alternative hypothesis is true, then by Corollary 2 of Székely et al. (2007), it follows that

$$T_n/n = \zeta_n = \frac{\mathcal{V}_n^2(X_i, Y_j)}{S_2(X_i, Y_j)} \to_P C > 0,$$

where C is a positive constant. For any sufficiently small number $0 < \epsilon < C$, we have

$$\lim_{n\to+\infty} P\{|\zeta_n - C| > \epsilon\} = 0.$$

Therefore,

$$\begin{aligned}
P_{ij}^{II}(n) &= P\big\{n\zeta_n \le \big(\Phi^{-1}(1-\alpha/2)\big)^2\big\} \\
&= P\big\{n\zeta_n \le \big(\Phi^{-1}(1-\alpha/2)\big)^2, |\zeta_n - C| > \epsilon\big\} + P\big\{n\zeta_n \le \big(\Phi^{-1}(1-\alpha/2)\big)^2, |\zeta_n - C| \le \epsilon\big\} \\
&\le P\{|\zeta_n - C| > \epsilon\} + P\Big\{\zeta_n - C + C \le \big(\Phi^{-1}(1-\alpha/2)\big)^2/n \;\Big|\; |\zeta_n - C| \le \epsilon\Big\} \\
&\le P\{|\zeta_n - C| > \epsilon\} + \mathbf{1}\big\{-\epsilon + C \le \big(\Phi^{-1}(1-\alpha/2)\big)^2/n\big\},
\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Since $-\epsilon + C > 0$, it remains to show that

$$\overline{\lim_{n\to\infty}} \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}} = 0.$$

Define $x_n = \Phi^{-1}(1-\alpha/2)$, then it is obvious that $x_n \to +\infty$. When n is sufficiently large, we have $1 - \Phi(x_n) \le 1/\sqrt{2\pi}\exp(-x_n^2/2)/x_n$ according to the property of the standard normal distribution. Therefore, when n is sufficiently large,

$$\alpha/2 = 1 - \Phi(x_n) \le \frac{1}{\sqrt{2\pi}x_n}\exp\left(-\frac{x_n^2}{2}\right),$$

i.e.,

$$\frac{x_n}{\sqrt{n}} \le \frac{2}{\sqrt{2\pi}}\frac{1}{n^{-3/2}}\exp\left(-\frac{x_n^2}{2}\right). \tag{S3.3}$$

If $\overline{\lim}_{n\to\infty} x_n/\sqrt{n} = \tau > 0$, then there exists a subsequence $\{x_{n_i}\}$ such that $\lim_{i\to+\infty} x_{n_i}/\sqrt{n_i} = \tau > 0$. For this subsequence, we have

$$\lim_{i\to+\infty} \frac{2}{\sqrt{2\pi}n_i^{-3/2}}\exp\left(-\frac{x_{n_i}^2}{2}\right) = 0.$$

This contradicts (S3.3), and thus

$$\lim_{n\to+\infty} P_{ij}^{II} \leq \overline{\lim_{n\to+\infty}} P\{|\zeta_n - C| > \epsilon\} + \mathbf{1}\big\{-\epsilon + C \leq \big(\Phi^{-1}(1-\alpha/2)\big)^2/n\big\} = 0.$$

Next we prove that $\hat{G}^+$ is consistent when $\alpha = O(n^{-2})$. Let $\mathbf{R} = (R_{ij})_{q\times p}$ where $R_{ij} = \mathbf{1}(X_i \not\perp\!\!\!\perp Y_j)$ and $\hat{\mathbf{R}}$ denote the estimate of $\mathbf{R}$ as shown in line 2 of Algorithm 1. According to Propositions S3.1 and S3.2, if $\hat{\mathbf{R}} = \mathbf{R}$ then $\hat{G}^+ = G^+$, which implies that

$$P(\hat{G}^+ \neq G^+) \leq P(\hat{\mathbf{R}} \neq \mathbf{R}).$$

Moreover,

$$P(\hat{\mathbf{R}} \neq \mathbf{R}) = P\bigg(\bigcup_{i=1}^{p}\{\hat{\mathbf{R}}_{i,+}(n) \neq \mathbf{R}_{i,+}\}\bigg) \leq \sum_{i=1}^{q}\sum_{j=1}^{p}\big\{P_{ij}^{I}(n) + P_{ij}^{II}(n)\big\}.$$

Therefore,

$$P(\hat{G}^+ \neq G^+) \leq P(\hat{\mathbf{R}} \neq \mathbf{R}) \leq \sum_{i=1}^{q}\sum_{j=1}^{p}\big\{P_{ij}^{I}(n) + P_{ij}^{II}(n)\big\}.$$

Since $\lim_{n\to+\infty} P_{ij}^{I}(n) = \lim_{n\to+\infty} P_{ij}^{II}(n) = 0$, it follows that

$$\lim_{n\to+\infty} P(\hat{G}^+ \neq G^+) \leq \sum_{i=1}^{q}\sum_{j=1}^{p}\bigg\{\lim_{n\to+\infty} P_{ij}^{I}(n) + \lim_{n\to+\infty} P_{ij}^{II}(n)\bigg\} = 0.$$

In conclusion, the estimated ARG is consistent, implying that the estimated candidate IV sets are also consistent. □

S3.3 Proof of Theorem 3

Proof. We first introduce some notations. Let $\mathbf{1}_d$ denote the d-dimensional vector with all elements equal to 1. For a vector $\mathbf{v}$, let $\mathbf{v} \odot \mathbf{w}$ denote the Hadamard product of it with another vector $\mathbf{w}$ of the same dimension. The diagonal matrix with diagonal elements being $\mathbf{v}$ is denoted as $\mathrm{diag}(\mathbf{v})$. To distinguish from $\mathbf{O}_i$ denoting the ith feature of a sample $\mathbf{O}$, we use $\mathbf{O}_{(i)}$ to denote the ith sample.

From Theorem 2, for any Borel set $B \subseteq \mathbb{R}^{|\hat{\mathcal{E}}|^+}$, we have

$$\begin{aligned}
&\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ \in B) \\
&= \lim_{n\to\infty} P(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ \in B \mid \hat{G}^+ = G^+) P(\hat{G}^+ = G^+) + P(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ \in B \mid \hat{G}^+ \neq G^+) P(\hat{G}^+ \neq G^+) \\
&= \lim_{n\to\infty} P(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ \in B \mid \hat{G}^+ = G^+),
\end{aligned}$$

where $\boldsymbol{\beta}^\circ = (\beta_{kj})_{(k,j)\in\hat{\mathcal{E}}^+}$. It is therefore sufficient to consider the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ only when $\hat{G}^+ = G^+$, and thus $\boldsymbol{\beta}^\circ = \boldsymbol{\beta}^*$ in such cases.

We first rewrite equation (4.1) in Algorithm 2 for further discussion. Following the same order of concatenating $\{M_{kj}(\boldsymbol{\beta})\}_{(k,j)\in\mathcal{E}^+}$ to obtain $\mathbf{M}(\boldsymbol{\beta})$ denoted by $(k_1, j_1), \ldots, (k_N, j_N)$, the basis functions $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$ of dimension $t(k;\gamma)$ can be arranged to

$$\mathbf{Z}_\gamma := \left(\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k_1)})^T,\ \ldots,\ \mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k_N)})^T\right)^T,$$

whose dimension is $t_\gamma = \sum_{j=1}^{p} \sum_{k\in \mathrm{an}_G(j)} t(k;\gamma)$. Let the vector $\mathbf{B}_\mathbf{Y} \in \mathbb{R}^{t_\gamma}$

denote $\big(Y_{j_1}\mathbf{1}^T_{t(k_1;\gamma)},\ldots,Y_{j_N}\mathbf{1}^T_{t(k_N;\gamma)}\big)^T$. Define $t(k_0;\gamma)=0$ and the matrix $\mathbf{A_Y}\in\mathbb{R}^{t_\gamma\times N}$ depending on $\mathbf{Y}$ with the sth row and lth column element given by:

$$(\mathbf{A_Y})_{sl}=\begin{cases} Y_{k_l} & ,\ k_l\in\{k_i\}\cup \mathrm{me}_G(k_i,j_i), j_l=j_i \\ 0 & ,\ \text{otherwise}\end{cases},\ \text{where}\ \sum_{\ell=0}^{i-1}t(k_\ell;\gamma)+1\le s\le\sum_{\ell=0}^{i}t(k_\ell;\gamma).$$

In other words, if the sth row of $\mathbf{B_Y}$ corresponds to the part $Y_{j_i}\mathbf{1}^T_{t(k_i;\gamma)}$, then the sth row of $\mathbf{A_Y}\in\mathbb{R}^{t_\gamma\times N}$ satisfies

$$(\mathbf{A_Y})^T_{s,\cdot}\boldsymbol{\beta}=\sum_{l=1}^{N}(\mathbf{A_Y})_{sl}\beta_l=\sum_{\ell\in\mathrm{me}_G(k_i,j_i)}\beta_{\ell j_i}Y_\ell+\beta_{k_ij_i}Y_{k_i},$$

for any $\boldsymbol{\beta}=(\beta_{k_1j_1},\ldots,\beta_{k_Nj_N})^T$. Since $\mathbf{Z}_\gamma$ is associated with $\boldsymbol{\mu}^*=E(\mathbf{X})$, let $\hat{\mathbf{Z}}_\gamma$ denote $\mathbf{Z}_\gamma$ with $\boldsymbol{\mu}^*$ substituted by its estimate $\hat{\boldsymbol{\mu}}$. This allows us to express (4.1) as the following equivalent problem:

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\hat{E}_n\big\{\hat{\mathbf{Z}}_\gamma\odot(\mathbf{B_Y}-\mathbf{A_Y}\boldsymbol{\beta})\big\}^T\boldsymbol{\Omega}\hat{E}_n\big\{\hat{\mathbf{Z}}_\gamma\odot(\mathbf{B_Y}-\mathbf{A_Y}\boldsymbol{\beta})\big\}. \quad \text{(S3.4)}$$

Therefore,

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\beta}}\hat{E}_n\big\{\hat{\mathbf{Z}}_\gamma\odot(\mathbf{B_Y}-\mathbf{A_Y}\boldsymbol{\beta})\big\}^T\boldsymbol{\Omega}\hat{E}_n\big\{\hat{\mathbf{Z}}_\gamma\odot(\mathbf{B_Y}-\mathbf{A_Y}\boldsymbol{\beta})\big\}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\\ &=-2\hat{E}_n(\mathrm{diag}(\hat{\mathbf{Z}}_\gamma)\mathbf{A_Y})^T\boldsymbol{\Omega}\hat{E}_n\big\{\mathrm{diag}(\hat{\mathbf{Z}}_\gamma)(\mathbf{B_Y}-\mathbf{A_Y}\hat{\boldsymbol{\beta}})\big\}\\ &=\mathbf{0}.\end{aligned}$$

By concatenating the estimation equations for $\boldsymbol{\mu}^*$ and $\boldsymbol{\beta}^*$, we have

$$\begin{aligned}\begin{bmatrix}\hat{\boldsymbol{\mu}}\\ \hat{\boldsymbol{\beta}}\end{bmatrix} &= \arg\min_{\boldsymbol{\mu},\boldsymbol{\beta}} \hat{E}_n \begin{bmatrix}\boldsymbol{\mu}-\mathbf{X}\\ \operatorname{diag}(\hat{\mathbf{Z}}_\gamma)(\mathbf{B_Y}-\mathbf{A_Y}\boldsymbol{\beta})\end{bmatrix}^T \cdot \begin{bmatrix}\mathbf{I}_q & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Omega}\hat{E}_n(\operatorname{diag}(\hat{\mathbf{Z}}_\gamma)\mathbf{A_Y})\hat{E}_n(\operatorname{diag}(\hat{\mathbf{Z}}_\gamma)\mathbf{A_Y})^T\boldsymbol{\Omega}\end{bmatrix}\\ &\quad \cdot \hat{E}_n \begin{bmatrix}\boldsymbol{\mu}-\mathbf{X}\\ \operatorname{diag}(\hat{\mathbf{Z}}_\gamma)(\mathbf{B_Y}-\mathbf{A_Y}\boldsymbol{\beta})\end{bmatrix}\\ &= \arg\min_{\boldsymbol{\mu},\boldsymbol{\beta}} \hat{E}_n \begin{bmatrix}\boldsymbol{\mu}-\mathbf{X}\\ \operatorname{diag}(\hat{\mathbf{Z}}_\gamma)(\mathbf{B_Y}-\mathbf{A_Y}\boldsymbol{\beta})\end{bmatrix}^T \mathbf{W}_n \hat{E}_n \begin{bmatrix}\boldsymbol{\mu}-\mathbf{X}\\ \operatorname{diag}(\hat{\mathbf{Z}}_\gamma)(\mathbf{B_Y}-\mathbf{A_Y}\boldsymbol{\beta})\end{bmatrix},\end{aligned}$$

where $\mathbf{W}_n$ is a data-adaptive positive semi-definite weighting matrix. For ease of notation, define $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}^T, \hat{\boldsymbol{\beta}}^T)^T$, $\boldsymbol{\theta}^* = (\boldsymbol{\mu}^{*T}, \boldsymbol{\beta}^{*T})^T$, and

$$\mathbf{m}_i = \mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta}) := \begin{bmatrix}\boldsymbol{\mu}-\mathbf{X}_{(i)}\\ \operatorname{diag}(\hat{\mathbf{Z}}_\gamma)_{(i)}(\mathbf{B}_{\mathbf{Y}(i)}-\mathbf{A}_{\mathbf{Y}(i)}\boldsymbol{\beta})\end{bmatrix}, \ for\ i = 1, \ldots, n.$$

Let $\mathcal{O} \subseteq \mathbb{R}^{p+q}$ denote the sample space of $\mathbf{O}_{(i)}$, and $\Theta \subseteq \mathbb{R}^{q+|\mathcal{E}|^+}$ the parameter space of $\boldsymbol{\theta}$, then $\mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})$ is a mapping from $\mathcal{O} \times \Theta$ to $\mathbb{R}^{q+t_\gamma}$. According to Theorem 3.2 of Hall (2005), we introduce the following regular conditions:

Assumption S3.1 (Regularity conditions on $\mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})$)**.** The function $\mathbf{m}_i : \mathcal{O} \times \Theta \to \mathbb{R}^{q+t_\gamma}$ satisfies that (i) it is continuous on Θ for each $\mathbf{O}_{(i)} \in \mathcal{O}$; (ii) $E(\mathbf{m}_i)$ is continuous on Θ.

Assumption S3.2 (Regularity conditions on $\partial\mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$)**.** (i) The

derivative matrix $\partial \mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$ exists and is continuous on Θ for each $\mathbf{O}_{(i)} \in \mathcal{O}$; (ii) $\boldsymbol{\theta}^*$ is an interior point of Θ; (iii) $E(\partial \mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^T)$exists and is finite.

Assumption S3.3 (Properties of $\boldsymbol{\Omega}$)**.** The weighting matrix $\boldsymbol{\Omega}$ may depend on data, but converges in probability to a positive definite matrix of constants.

Under Assumption 4, when the positive semi-definite matrix $\boldsymbol{\Omega}$ satisfies the regularity condition S3.3, $\mathbf{W}_n$ convergence in probability to a positive definite matrix $\mathbf{W}_{\boldsymbol{\Omega}}$ as $n \to \infty$,

$$\mathbf{W}_n \to \mathbf{W}_{\boldsymbol{\Omega}} = \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega} E(\mathrm{diag}(\mathbf{Z}_\gamma)\mathbf{A}_\mathbf{Y}) E(\mathrm{diag}(\mathbf{Z}_\gamma)\mathbf{A}_\mathbf{Y})^T \boldsymbol{\Omega} \end{bmatrix}.$$

Assumption S3.4 (Compactness of Θ)**.** The parameter space Θ is a compact set.

Assumption S3.5 (Domination of $\mathbf{m}(\mathbf{O}_{(i)}; \boldsymbol{\theta})$)**.** $E\{\sup_{\boldsymbol{\theta} \in \Theta} \|\mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})\|_2\} < \infty$.

Assumption S3.6 (Properties of the variance of the sample moment)**.** (i) The moment $E\{\mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta}^*)\mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta}^*)^T\}$ exists and is finite; (ii) The limit

$$\lim_{n \to +\infty} \mathrm{var}\left(\sqrt{n} \hat{E}_n \begin{bmatrix} \boldsymbol{\mu}^* - \mathbf{X} \\ \mathrm{diag}(\mathbf{Z}_\gamma)(\mathbf{B}_\mathbf{Y} - \mathbf{A}_\mathbf{Y} \boldsymbol{\beta}^*) \end{bmatrix} \right)$$

exists and is finite.

Assumption S3.7 (Continuity of $E\{\partial \mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^T\}$)**.** The function $E\{\partial \mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^T\}$ is continuous on some neighborhood N_ϵ of $\boldsymbol{\theta}^*$.

Assumption S3.8 (Uniform Convergence of $\mathbf{G}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \partial \mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$)**.**

$$\sup_{\boldsymbol{\theta}\in N_\epsilon} \left\| \mathbf{G}_n(\boldsymbol{\theta}) - E\{\partial \mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^T\} \right\|_2 \to_P 0.$$

Suppose that Assumptions 1–4 and the regularity conditions S3.1–S3.7 hold. According to Theorem 3.2 of Hall (2005), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} N\left(\mathbf{0}, (\mathbf{G}^T\mathbf{W}_{\boldsymbol{\Omega}}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}_{\boldsymbol{\Omega}}\mathbf{F}\mathbf{W}_{\boldsymbol{\Omega}}\mathbf{G}(\mathbf{G}^T\mathbf{W}_{\boldsymbol{\Omega}}\mathbf{G})^{-1}\right),$$

where

$$\mathbf{G} = E\left\{ \left. \frac{\partial \mathbf{m}(\mathbf{O}_{(i)}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right\} = \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{C} & \mathbf{D} \end{bmatrix},$$

$$\mathbf{F} = \lim_{n\to+\infty} \operatorname{var}\left\{ \sqrt{n}\hat{E}_n \begin{bmatrix} \boldsymbol{\mu}^* - \mathbf{X} \\ \operatorname{diag}(\mathbf{Z}_\gamma)(\mathbf{B}_\mathbf{Y} - \mathbf{A}_\mathbf{Y}\boldsymbol{\beta}^*) \end{bmatrix} \right\},$$

with

$$\mathbf{C} = E\left\{ \left. \frac{\partial}{\partial \boldsymbol{\mu}} \operatorname{diag}(\hat{\mathbf{Z}}_\gamma)(\mathbf{B}_\mathbf{Y} - \mathbf{A}_\mathbf{Y}\boldsymbol{\beta}) \right|_{\boldsymbol{\mu}=\boldsymbol{\mu}^*, \boldsymbol{\beta}=\boldsymbol{\beta}^*} \right\},$$

$$\mathbf{D} = E\left\{ \left. \frac{\partial}{\partial \boldsymbol{\beta}} \operatorname{diag}(\hat{\mathbf{Z}}_\gamma)(\mathbf{B}_\mathbf{Y} - \mathbf{A}_\mathbf{Y}\boldsymbol{\beta}) \right|_{\boldsymbol{\mu}=\boldsymbol{\mu}^*, \boldsymbol{\beta}=\boldsymbol{\beta}^*} \right\} = -E\{\operatorname{diag}(\mathbf{Z}_\gamma)\mathbf{A}_\mathbf{Y}\}.$$

In particular, the matrix $\mathbf{C}$ depends on the form of $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$. For exam-

ple, when the candidate IVs all take values in $\{0, 1\}$,

$$\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)}) = \big(\Pi_{s\in\alpha_k(1)}(X_s - \hat{\mu}_s), \ldots, \Pi_{s\in\alpha_k(t(k;\gamma))}(X_s - \hat{\mu}_s)\big)^T,$$

where $\alpha_k(\cdot)$ are the elements of $\{\alpha : \alpha \subseteq \mathrm{ca}_G(k),\ |\alpha| \geq |\,\mathrm{ca}_G(k)| - \gamma + 1\}$ as described in the main text. Therefore for each $j = 1, \ldots, q$, the partial derivation is

$$\left.\frac{\partial \mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})}{\partial \mu_j}\right|_{\boldsymbol{\mu}=\boldsymbol{\mu}^*,\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \left[\left\{-\mathbf{1}(j \in \alpha_k(\ell)) \prod_{s\in\alpha_k(\ell), s\neq j} (X_s - \mu_s^*)\right\}_{\ell=1}^{t(k;\gamma)}\right]^T.$$

Since we are only interested in the asymptotic properties of $\hat{\boldsymbol{\beta}}$, Theorem 3 holds by extracting the corresponding submatrix. □

S3.4 Proof of Theorem 4

Proof. Under Assumptions 1–3, it has been shown that $\hat{G}^+$ is consistent according to Theorem 2. Moreover, as long as $\hat{G}^+ = G^+$, we have $\mathrm{RE}(n) = 0$. Therefore,

$$\begin{aligned}
&\lim_{n\to\infty} E\{\mathrm{RE}(n)\} \\
&\quad= \lim_{n\to\infty} E\{\mathrm{RE}(n) \mid \hat{G}^+ \neq G^+\} P\big(\hat{G}^+ \neq G^+\big) + \lim_{n\to\infty} E\{\mathrm{RE}(n) \mid \hat{G}^+ = G^+\} P(\hat{G}^+ = G^+) \\
&\quad\leq |\mathcal{E}| \lim_{n\to\infty} P(\hat{G}^+ \neq G^+) + 0 \\
&\quad= 0.
\end{aligned}$$

Note that q^* is the target FDR level in Algorithm 2. Applying Theorem 1.3 of Benjamini and Yekutieli (2001), we have

$$\lim_{n\to\infty} E\left\{ \frac{\mathrm{FP}(n)}{\mathrm{TP}(n)+\mathrm{RE}(n)+\mathrm{FP}(n)} \,\middle|\, \hat{G}^+ = G^+ \right\} \leq q^*.$$

Since the $\hat{G}^+$ is consistent and $\mathcal{E} \neq \emptyset$, it follows that

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{FDR}(\hat{\mathcal{E}}) \\
&\quad = \lim_{n\to\infty} E\left\{ \frac{\mathrm{RE}(n)}{\mathrm{TP}(n)+\mathrm{RE}(n)+\mathrm{FP}(n)} \right\} + \lim_{n\to\infty} E\left\{ \frac{\mathrm{FP}(n)}{\mathrm{TP}(n)+\mathrm{RE}(n)+\mathrm{FP}(n)} \right\} \\
&\quad = 0 + \lim_{n\to\infty} E\left\{ \frac{\mathrm{FP}(n)}{\mathrm{TP}(n)+\mathrm{RE}(n)+\mathrm{FP}(n)} \,\middle|\, \hat{G}^+ = G^+ \right\} P(\hat{G}^+ = G^+) \\
&\quad + \lim_{n\to\infty} E\left\{ \frac{\mathrm{FP}(n)}{\mathrm{TP}(n)+\mathrm{RE}(n)+\mathrm{FP}(n)} \,\middle|\, \hat{G}^+ \neq G^+ \right\} P(\hat{G}^+ \neq G^+) \\
&\quad = \lim_{n\to\infty} E\left\{ \frac{\mathrm{FP}(n)}{\mathrm{TP}(n)+\mathrm{RE}(n)+\mathrm{FP}(n)} \,\middle|\, \hat{G}^+ = G^+ \right\} \\
&\quad \leq q^*,
\end{aligned}$$

which completes the proof. □

S4 Simulation details and extended results

In this section, we present the implementation details and extended results of the simulation studies. Computationally, we use the R package `grivet` for the implementation of GrIVET, `pcalg` for RFCI, and `lrpsadmm` and `pcalg` for LRpS-GES.

We consider two types of DAGs with unobserved confounders: random

graphs and hub graphs. Let $\mathbf{A} \in \mathbb{R}^{p\times p}$ denote the adjacency matrix for the DAG. For random graphs, the upper off-diagonal entries of $\mathbf{A}$ are independently sampled from Bernoulli$(1/(2p))$, while the other entries are set to 0. For hub graphs, the entries A_{1j}, $j = 2, \ldots, p$, are set to 1, with the remaining set to 0. Further, if $A_{kj} \neq 0$, then β^*_{kj} is sampled from the uniform distribution on $(-1.2, -0.8) \cup (0.8, 1.2)$; otherwise, it is set to 0. We consider the SEM

$$Y_j = \sum_{i=1}^{p} \beta^*_{ij} Y_i + g_j(\mathbf{X}_{\mathrm{in}_G(j)}) + \boldsymbol{\phi}_j^T \mathbf{U} + e_j, \quad j = 1, \ldots, p, \tag{S4.1}$$

where $\mathbf{e} = (e_1, \ldots, e_p)^T \sim N_p(\mathbf{0}, \boldsymbol{\Lambda})$ and the unobserved confounders $\mathbf{U} \sim N_r(\mathbf{0}, \mathbf{I}_r)$. Here $\boldsymbol{\Lambda} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with σ_i sampled uniformly from $(0.3, 0.4)$. The coefficients $\boldsymbol{\phi}_j = (\phi_{1j}, \ldots, \phi_{rj})^T$ are set as follows: ϕ_{11} and ϕ_{kj}, $j = 2k-1, 2k$, $k = 1, \ldots, r$, are sampled uniformly from $(-0.4, -0.3) \cup (0.3, 0.4)$, while the other entries are set to 0. We set $q = 2p + \lfloor p/2 \rfloor$ and $\mathrm{in}_G(j) = \{j, p+j, 2p+\lfloor j/2 \rfloor\}$. Hence, each X_ℓ, $\ell = 2p, \ldots, 2p+\lfloor (p-1)/2 \rfloor$, intervenes on two primary variables, while any other X_ℓ intervenes on a single Y_j.

We consider two types of secondary variables $\mathbf{X}$: (1) the continuous case where $X_i \sim N(0, 1)$ independently, and (2) the discrete case where $X_i \sim$ Bernoulli(0.5) independently. Depending on the types of $\mathbf{X}$, we specify

$g_j(\mathbf{X}_{\mathrm{in}_G(j)})$ as follows: for the continuous case,

$$g_j(\mathbf{X}_{\mathrm{in}_G(j)}) = C\left(w_j \sum_{\ell \in \mathrm{in}_G(j)} \{X_\ell^2 + \mathbf{1}(X_\ell > 0)\} + \frac{w_j}{2} \sum_{k,\ell \in \mathrm{in}_G(j),\, k \neq \ell} X_k X_\ell\right), \tag{S4.2}$$

where $C = 1$ and w_j are sampled uniformly from $(-3.2, -2.8) \cup (2.8, 3.2)$; for the discrete case,

$$g_j(\mathbf{X}_{\mathrm{in}_G(j)}) = C\left(\sum_{k,\ell \in \mathrm{in}_G(j),\, k \neq \ell} X_k X_\ell\right). \tag{S4.3}$$

To examine our method for DAGs of different sizes, we fix the sample size at $n = 1000$ while varying the dimension as $(p, q, r) = (10, 25, 5)$ and $(20, 50, 10)$. We set $\gamma = 1$ for PLACID. In Algorithm 2, we set the weighting matrix $\mathbf{\Omega} = \mathbf{I}$ and the FDR level $q^* = 0.05$. For the continuous case, we use tensor products of polynomial bases to approximate $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$. All simulations are repeated 100 times.

To empirically measure the strength of surrogate IVs, we present the average first-stage F-statistics in Table S4.1. All values are well above the commonly used threshold of 10 (Staiger and Stock, 1997), suggesting that the surrogate IVs are sufficiently strong across all simulation settings.

To complement the simulation results in Section 5, where we set $\gamma = 1$ to construct surrogate IVs, we conduct additional simulations with $\gamma = 2 = |\,\mathrm{iv}_G(j)|$ for all $j = 1, \ldots, p$. Tables S4.2 and S4.3 present the performance of PLACID under this setting. The corresponding average first-stage F-

Table S4.1: Means of the average first-stage F-statistic as an empirical measure of surrogate IV strength for PLACID when $C = 1$ and $\gamma = 1$.

Setting	Graph	p	F-statistic
Continuous	Random	10	29.48
		20	28.92
	Hub	10	25.84
		20	23.35
Discrete	Random	10	56.74
		20	58.26
	Hub	10	49.47
		20	45.19

statistics are reported in Table S4.4. For continuous secondary variables, the results in Tables S4.2 and S4.3 are comparable to those in Tables 1 and 3, which indicates that PLACID appears insensitive to the choice of γ under certain conditions. This can be attributed to the comparable strength of the surrogate IVs induced by $\gamma = 1$ and $\gamma = 2$, as evidenced by the similar average first-stage F-statistics for continuous secondary variables in Tables S4.1 and S4.4. For discrete secondary variables, PLACID with $\gamma = 2$ performs slightly better for parameter estimation, which may be due to the stronger surrogate IVs induced by this larger value of γ.

Table S4.2: Means and standard deviations (in parentheses) of different causal discovery performance metrics for PLACID with continuous and discrete secondary variables when $C = 1$ and $\gamma = 2$.

Setting	Graph	p	FDP	TPR	SHD	JI
Continuous	Random	10	0.02(0.08)	0.95(0.14)	0.26(0.61)	0.93(0.16)
		20	0.02(0.08)	0.92(0.17)	0.43(0.87)	0.91(0.18)
	Hub	10	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		20	0.00(0.00)	1.00(0.01)	0.03(0.17)	1.00(0.01)
Discrete	Random	10	0.01(0.04)	0.95(0.19)	0.10(0.41)	0.95(0.19)
		20	0.01(0.03)	0.99(0.05)	0.18(0.54)	0.98(0.06)
	Hub	10	0.00(0.00)	0.95(0.09)	0.41(0.79)	0.95(0.09)
		20	0.00(0.00)	0.96(0.05)	0.76(1.04)	0.96(0.05)

Table S4.3: Means and standard deviations (in parentheses) of different estimation losses for PLACID with continuous and discrete secondary variables when $C = 1$ and $\gamma = 2$.

Setting	Graph	p	L_∞	L_1	L_2
Continuous	Random	10	0.19(0.36)	0.30(0.63)	0.22(0.43)
		20	0.38(0.39)	0.78(0.90)	0.47(0.49)
	Hub	10	0.07(0.02)	0.28(0.08)	0.11(0.03)
		20	0.12(0.06)	0.82(0.26)	0.24(0.08)
Discrete	Random	10	0.10(0.24)	0.16(0.39)	0.12(0.28)
		20	0.17(0.37)	0.36(0.79)	0.22(0.49)
	Hub	10	0.30(0.36)	0.78(0.78)	0.40(0.45)
		20	0.44(0.41)	1.60(1.11)	0.64(0.53)

Table S4.4: Means of the average first-stage F-statistic as an empirical measure of surrogate IV strength for PLACID when $C = 1$ and $\gamma = 2$.

Setting	Graph	p	F-statistic
Continuous	Random	10	34.36
		20	27.83
	Hub	10	21.47
		20	20.75
Discrete	Random	10	99.11
		20	96.83
	Hub	10	90.18
		20	86.33

S5 Additional simulation studies

In this section, we provide additional simulation studies. Results with correlated secondary variables are presented in Sections S5.1, and results with varying IV strengths, sample sizes, values of γ, and numbers of valid IVs are given in Sections S5.2, S5.3, S5.4, and S5.5, respectively. All simulations are repeated 100 times.

S5.1 Correlated secondary variables

In this subsection, we conduct additional simulations to examine how PLACID performs when Assumption 1 is violated.

We consider the same settings as in Section S4 with $n = 1000$, $(p, q, r) = (10, 25, 5)$, $C = 1$, and $\gamma = 1$, but modify the data-generating mechanism for $\mathbf{X}$ to induce dependence. For the continuous case, we let $\mathbf{X}$ follow a mean-zero multivariate Gaussian distribution whose (i, j)-th correlation is $0.1^{|i-j|}$. For the discrete case, we first generate a latent Gaussian vector $\widetilde{\mathbf{X}}$ from the same distribution and then define $X_i = \mathbf{1}(\widetilde{X}_i > 0)$ for $i = 1, \ldots, q$.

The causal discovery and parameter estimation results are reported in Tables S5.1 and S5.2, respectively. We also report the empirical measures of surrogate IV strength in Table S5.3. Comparisons of Tables 1, 2, and S5.1 for causal discovery, as well as Tables 3 and S5.2 for parameter estimation, show that the accuracy of PLACID deteriorates when Assumption 1 is violated, as theoretically expected. Such a decline is especially apparent for continuous secondary variables. Nevertheless, PLACID remains superior to the competing methods across all scenarios, demonstrating its robustness.

Table S5.1: Means and standard deviations (in parentheses) of different causal discovery performance metrics for four methods with correlated secondary variables.

Setting	Graph	Method	FDP	TPR	SHD	JI
Continuous	Random	PLACID	0.03(0.10)	0.87(0.24)	0.48(0.90)	0.86(0.25)
		GrIVET	0.50(0.38)	0.50(0.38)	3.07(2.43)	0.31(0.31)
		RFCI	0.00(0.05)	0.63(0.36)	1.04(1.09)	0.63(0.36)
		LRpS-GES	0.64(0.17)	0.99(0.04)	3.87(0.79)	0.35(0.17)
	Hub	PLACID	0.00(0.00)	0.89(0.24)	0.98(2.20)	0.89(0.24)
		GrIVET	0.38(0.37)	0.44(0.37)	7.28(4.81)	0.38(0.34)
		RFCI	0.01(0.04)	0.57(0.24)	3.96(2.19)	0.56(0.24)
		LRpS-GES	0.38(0.05)	0.85(0.11)	6.08(1.46)	0.56(0.09)
Discrete	Random	PLACID	0.02(0.08)	0.91(0.22)	0.28(0.57)	0.89(0.22)
		GrIVET	0.19(0.31)	0.62(0.38)	1.35(1.28)	0.53(0.37)
		RFCI	0.00(0.00)	0.80(0.30)	0.43(0.64)	0.80(0.30)
		LRpS-GES	0.76(0.15)	0.97(0.13)	6.31(1.48)	0.24(0.15)
	Hub	PLACID	0.00(0.00)	0.95(0.09)	0.48(0.83)	0.95(0.09)
		GrIVET	0.00(0.02)	0.50(0.15)	4.47(1.36)	0.50(0.15)
		RFCI	0.00(0.00)	0.38(0.19)	5.59(1.71)	0.38(0.19)
		LRpS-GES	0.47(0.05)	0.77(0.06)	8.25(1.23)	0.46(0.05)

Table S5.2: Means and standard deviations (in parentheses) of different estimation losses for two methods with correlated secondary variables.

Setting	Graph	Method	L_∞	L_1	L_2
Continuous	Random	PLACID	0.44(0.43)	0.77(0.99)	0.52(0.55)
		GrIVET	0.91(0.36)	2.15(1.59)	1.22(0.60)
	Hub	PLACID	0.49(0.37)	2.14(2.24)	0.85(0.79)
		GrIVET	1.02(0.27)	6.96(3.64)	2.31(0.92)
Discrete	Random	PLACID	0.25(0.36)	0.37(0.52)	0.28(0.39)
		GrIVET	0.83(0.37)	1.59(1.11)	1.04(0.53)
	Hub	PLACID	0.41(0.36)	1.04(0.71)	0.53(0.42)
		GrIVET	1.16(0.04)	8.96(0.37)	3.00(0.12)

Table S5.3: Means of the average first-stage F-statistic as an empirical measure of surrogate IV strength for PLACID with correlated secondary variables.

Setting	Graph	F-statistic
Continuous	Random	33.73
	Hub	26.34
Discrete	Random	57.21
	Hub	49.61

S5.2 Varying IV strengths

In this subsection, we conduct additional simulations to examine the performance of PLACID under varying surrogate IV strengths.

We consider the same settings as in Section S4 with $n = 1000$, $(p, q, r) = (10, 25, 5)$, and $\gamma = 1$. To vary the IV strength, we set the coefficient C to either 0.6 or 1.5 in the data-generating models (S4.2) and (S4.3). The empirical measures of surrogate IV strength are summarized in Table S5.4. As expected, Table S5.4 shows that IV strength increases with larger values of C.

Table S5.4: Means of the average first-stage F-statistic as an empirical measure of surrogate IV strength for PLACID with varying C when $\gamma = 1$.

Setting	Graph	C	F-statistic
Continuous	Random	0.6	16.01
		1.5	47.58
	Hub	0.6	18.75
		1.5	39.12
Discrete	Random	0.6	19.05
		1.5	71.83
	Hub	0.6	18.82
		1.5	65.76

The causal discovery results for the continuous and discrete cases are presented in Tables S5.5 and S5.6, respectively, with parameter estimation results in Table S5.7. Overall, these results indicate that stronger surrogate IVs lead to better performance of PLACID in both causal discovery and parameter estimation. While performance degrades under weak IV settings as expected, PLACID consistently outperforms the competing methods across all scenarios. In addition, it effectively controls the FDP below the nominal level $q^* = 0.05$, demonstrating its robustness.

Table S5.5: Means and standard deviations (in parentheses) of different causal discovery performance metrics for four methods with continuous secondary variables across varying surrogate IV strengths.

Graph	C	Method	FDP	TPR	SHD	JI
Random	0.6	PLACID	0.02(0.09)	0.86(0.24)	0.51(0.94)	0.85(0.25)
		GrIVET	0.11(0.26)	0.15(0.32)	2.16(1.59)	0.12(0.26)
		RFCI	0.02(0.12)	0.41(0.38)	1.20(1.06)	0.41(0.38)
		LRpS-GES	0.68(0.17)	0.87(0.32)	4.41(1.06)	0.31(0.17)
	1.5	PLACID	0.01(0.04)	0.95(0.17)	0.15(0.44)	0.94(0.17)
		GrIVET	0.59(0.38)	0.42(0.37)	3.76(2.76)	0.25(0.27)
		RFCI	0.01(0.07)	0.50(0.40)	0.94(0.94)	0.50(0.40)
		LRpS-GES	0.68(0.17)	0.86(0.33)	4.10(0.97)	0.32(0.17)
Hub	0.6	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.16(0.25)	0.35(0.40)	6.82(3.07)	0.30(0.34)
		RFCI	0.01(0.04)	0.63(0.21)	3.34(1.90)	0.63(0.21)
		LRpS-GES	0.42(0.04)	0.79(0.07)	7.09(1.09)	0.50(0.06)
	1.5	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.33(0.32)	0.51(0.36)	6.59(4.30)	0.43(0.33)
		RFCI	0.01(0.04)	0.58(0.21)	3.80(1.86)	0.58(0.21)
		LRpS-GES	0.40(0.05)	0.82(0.09)	6.54(1.31)	0.53(0.08)

Table S5.6: Means and standard deviations (in parentheses) of different causal discovery performance metrics for four methods with discrete secondary variables across varying surrogate IV strengths.

Graph	C	Method	FDP	TPR	SHD	JI
Random	0.6	PLACID	0.02(0.10)	0.83(0.28)	0.54(1.06)	0.82(0.28)
		GrIVET	0.07(0.19)	0.59(0.39)	1.06(1.09)	0.55(0.37)
		RFCI	0.00(0.00)	0.70(0.34)	0.60(0.83)	0.70(0.34)
		LRpS-GES	0.79(0.14)	0.79(0.36)	7.91(1.71)	0.20(0.14)
	1.5	PLACID	0.00(0.00)	0.97(0.10)	0.09(0.29)	0.97(0.10)
		GrIVET	0.24(0.33)	0.60(0.37)	1.67(1.76)	0.50(0.35)
		RFCI	0.00(0.00)	0.89(0.26)	0.25(0.69)	0.89(0.26)
		LRpS-GES	0.67(0.15)	0.93(0.23)	4.29(1.33)	0.33(0.15)
Hub	0.6	PLACID	0.00(0.00)	0.79(0.20)	1.88(1.83)	0.79(0.20)
		GrIVET	0.00(0.00)	0.50(0.17)	4.50(1.54)	0.50(0.17)
		RFCI	0.00(0.00)	0.36(0.20)	5.75(1.80)	0.36(0.20)
		LRpS-GES	0.50(0.08)	0.76(0.08)	9.21(2.02)	0.43(0.09)
	1.5	PLACID	0.00(0.00)	1.00(0.02)	0.04(0.20)	1.00(0.02)
		GrIVET	0.00(0.00)	0.49(0.17)	4.57(1.57)	0.49(0.17)
		RFCI	0.00(0.00)	0.50(0.21)	4.51(1.89)	0.50(0.21)
		LRpS-GES	0.40(0.05)	0.82(0.09)	6.56(1.28)	0.53(0.08)

Table S5.7: Means and standard deviations (in parentheses) of different estimation losses for two methods with continuous and discrete secondary variables across varying surrogate IV strengths.

Setting	Graph	C	Method	L_∞	L_1	L_2
Continuous	Random	0.6	PLACID	0.28(0.35)	0.46(0.73)	0.32(0.44)
			GrIVET	0.94(0.35)	2.20(1.51)	1.34(0.63)
		1.5	PLACID	0.23(0.33)	0.36(0.57)	0.26(0.38)
			GrIVET	0.23(0.33)	1.94(1.39)	1.09(0.56)
	Hub	0.6	PLACID	0.18(0.06)	0.69(0.22)	0.28(0.09)
			GrIVET	1.16(0.04)	9.03(0.37)	3.02(0.12)
		1.5	PLACID	0.08(0.03)	0.32(0.08)	0.13(0.03)
			GrIVET	0.99(0.28)	6.29(3.56)	2.15(0.92)
Discrete	Random	0.6	PLACID	0.41(0.46)	0.70(1.10)	0.49(0.59)
			GrIVET	0.81(0.37)	1.73(1.29)	1.08(0.60)
		1.5	PLACID	0.11(0.22)	0.17(0.36)	0.13(0.26)
			GrIVET	0.82(0.37)	1.90(1.61)	1.10(0.62)
	Hub	0.6	PLACID	0.81(0.35)	2.38(1.63)	1.17(0.65)
			GrIVET	1.17(0.03)	9.04(0.34)	3.03(0.11)
		1.5	PLACID	0.16(0.14)	0.53(0.21)	0.23(0.14)
			GrIVET	1.13(0.07)	6.79(1.35)	2.43(0.36)

S5.3 Varying sample sizes

In this subsection, we conduct additional simulations to assess the performance of PLACID under varying sample sizes.

We consider the same settings as in Section S4 with $(p, q, r) = (10, 25, 5)$, $C = 1$, and $\gamma = 1$. In addition to the moderate sample size $n = 1000$ studied in Section 5 or Section S4, we examine two additional scenarios with $n = 500$ and $n = 2000$.

The causal discovery results for the continuous and discrete cases are presented in Tables S5.8 and S5.9, respectively, with parameter estimation results in Table S5.10. We also report the empirical measures of surrogate IV strength in Table S5.11. Comparisons between Tables 1–3 and S5.8–S5.10 show that the accuracy of PLACID improves with increasing sample size across all settings. For causal discovery, the FDP is well controlled at or below the nominal level $q^* = 0.05$ across all sample sizes. For parameter estimation, the estimation error decreases as the sample size increases, confirming the consistency of PLACID. Moreover, PLACID outperforms the competing methods across all scenarios, including the small-sample case with $n = 500$, demonstrating its robustness.

Table S5.8: Means and standard deviations (in parentheses) of different causal discovery performance metrics for four methods with continuous secondary variables across varying sample sizes.

Setting	n	Method	FDP	TPR	SHD	JI
Random	500	PLACID	0.05(0.15)	0.79(0.31)	0.69(1.18)	0.77(0.32)
		GrIVET	0.75(0.31)	0.33(0.37)	4.85(3.07)	0.17(0.23)
		RFCI	0.01(0.10)	0.66(0.39)	0.76(0.93)	0.66(0.39)
		LRpS-GES	0.66(0.16)	0.96(0.18)	3.94(0.79)	0.34(0.16)
	2000	PLACID	0.01(0.05)	0.97(0.12)	0.15(0.44)	0.96(0.13)
		GrIVET	0.37(0.39)	0.63(0.39)	2.02(1.99)	0.46(0.36)
		RFCI	0.00(0.05)	0.71(0.36)	0.82(1.05)	0.71(0.36)
		LRpS-GES	0.70(0.16)	0.98(0.08)	4.30(1.07)	0.30(0.15)
Hub	500	PLACID	0.00(0.00)	0.98(0.07)	0.18(0.67)	0.98(0.07)
		GrIVET	0.55(0.33)	0.34(0.30)	10.19(5.29)	0.26(0.26)
		RFCI	0.06(0.11)	0.52(0.22)	4.65(2.06)	0.50(0.22)
		LRpS-GES	0.42(0.05)	0.77(0.11)	6.96(1.36)	0.50(0.09)
	2000	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.19(0.30)	0.57(0.39)	4.68(4.37)	0.55(0.39)
		RFCI	0.00(0.01)	0.64(0.22)	3.27(1.98)	0.64(0.22)
		LRpS-GES	0.38(0.05)	0.85(0.11)	6.02(1.45)	0.56(0.09)

Table S5.9: Means and standard deviations (in parentheses) of different causal discovery performance metrics for four methods with discrete secondary variables across varying sample sizes.

Setting	n	Method	FDP	TPR	SHD	JI
Random	500	PLACID	0.02(0.07)	0.76(0.33)	0.68(1.02)	0.75(0.33)
		GrIVET	0.16(0.29)	0.69(0.36)	1.15(1.12)	0.61(0.36)
		RFCI	0.00(0.00)	0.69(0.36)	0.74(0.82)	0.69(0.36)
		LRpS-GES	0.71(0.18)	0.96(0.15)	4.72(1.36)	0.28(0.18)
	2000	PLACID	0.01(0.04)	0.94(0.24)	0.06(0.34)	0.93(0.24)
		GrIVET	0.19(0.31)	0.65(0.33)	1.33(1.12)	0.55(0.34)
		RFCI	0.00(0.03)	0.74(0.36)	0.65(0.88)	0.73(0.36)
		LRpS-GES	0.78(0.14)	0.94(0.17)	7.51(1.44)	0.22(0.13)
Hub	500	PLACID	0.00(0.00)	0.75(0.21)	2.26(1.86)	0.75(0.21)
		GrIVET	0.01(0.04)	0.49(0.17)	4.63(1.50)	0.49(0.17)
		RFCI	0.00(0.00)	0.37(0.18)	5.66(1.63)	0.37(0.18)
		LRpS-GES	0.39(0.09)	0.76(0.09)	6.66(1.87)	0.51(0.10)
	2000	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.00(0.00)	0.50(0.17)	4.51(1.56)	0.50(0.17)
		RFCI	0.00(0.00)	0.47(0.18)	4.73(1.62)	0.47(0.18)
		LRpS-GES	0.49(0.05)	0.79(0.07)	8.85(1.30)	0.45(0.05)

Table S5.10: Means and standard deviations (in parentheses) of different estimation losses for two methods with continuous and discrete secondary variables across varying sample sizes.

Setting	Graph	n	Method	L_∞	L_1	L_2
Continuous	Random	500	PLACID	0.43(0.44)	0.86(1.25)	0.55(0.64)
			GrIVET	0.95(0.29)	2.49(1.40)	1.33(0.53)
		2000	PLACID	0.17(0.25)	0.27(0.41)	0.20(0.29)
			GrIVET	0.70(0.44)	1.28(1.03)	0.84(0.57)
	Hub	500	PLACID	0.30(0.28)	0.94(0.78)	0.42(0.33)
			GrIVET	1.12(0.18)	9.19(3.70)	2.78(0.76)
		2000	PLACID	0.08(0.02)	0.32(0.09)	0.13(0.03)
			GrIVET	0.84(0.41)	5.10(3.70)	1.81(1.11)
Discrete	Random	500	PLACID	0.49(0.46)	0.81(1.04)	0.58(0.60)
			GrIVET	0.79(0.39)	1.62(1.21)	1.03(0.58)
		2000	PLACID	0.04(0.14)	0.08(0.35)	0.05(0.37)
			GrIVET	0.87(0.35)	1.83(1.18)	1.13(0.53)
	Hub	500	PLACID	0.82(0.35)	2.67(1.62)	1.28(0.65)
			GrIVET	1.16(0.04)	9.00(0.36)	3.01(0.11)
		2000	PLACID	0.11(0.03)	0.43(0.12)	0.17(0.05)
			GrIVET	1.16(0.04)	9.04(0.35)	3.03(0.12)

Table S5.11: Means of the average first-stage F-statistic as an empirical measure of surrogate IV strength for PLACID across varying sample sizes with $C = 1$ and $\gamma = 1$.

Setting	Graph	n	F-statistic
Continuous	Random	500	19.43
		2000	62.91
	Hub	500	13.15
		2000	46.83
Discrete	Random	500	30.14
		2000	113.20
	Hub	500	24.53
		2000	99.17

S5.4 Varying values of γ

In this subsection, we conduct additional simulations to assess the performance of PLACID under different values of γ.

To further evaluate the scenarios in which $\gamma > |\mathrm{iv}_G(j)|$ for some j, so that Assumption 3 is violated, we consider a modified data-generating process that differs from the one described in Section S4. Specifically, we maintain most of the simulation setup in Section S4 while altering the settings of $\mathrm{in}_G(j)$ and $g_j(\mathbf{X}_{\mathrm{in}_G(j)})$ for $j = 1, \ldots, p$. We set $q = 3p$ and $\mathrm{ca}_G(j) = \{j, j+p, j+2p\}$ for all Y_j. For each X_ℓ that is a candidate IV of Y_j but not a valid IV of Y_j, we let X_ℓ have causal effects on all descendants of Y_j. That is, for every $\ell \in \mathrm{ca}_G(j) \setminus \mathrm{iv}_G(j)$, we set $\ell \in \mathrm{in}_G(k)$ if $j \in \mathrm{an}_G(k)$. According to Definitions 1 and 2, this setup enables us to control the number of valid IVs for Y_j by adjusting whether its candidate IVs have causal effects on its descendants.

Depending on the types of $\mathbf{X}$, we specify $g_j(\mathbf{X}_{\mathrm{in}_G(j)})$ as follows: for the continuous case,

$$g_j(\mathbf{X}_{\mathrm{in}_G(j)}) = w_j \Bigg(\sum_{\ell \in \mathrm{in}_G(j)} \{X_\ell^2 + \mathbf{1}(X_\ell > 0)\} + \sum_{k,\ell \in \mathrm{in}_G(j),\, k \neq \ell} X_k X_\ell + \sum_{i,k,\ell \in \mathrm{in}_G(j), i \neq k, k \neq \ell, \ell \neq i} X_i X_k X_\ell \Bigg),$$

where w_j are sampled uniformly from $(-3.2, -2.8) \cup (2.8, 3.2)$; for the discrete case,

$$g_j(\mathbf{X}_{\text{in}_G(j)}) = \sum_{k,\ell \in \text{in}_G(j),\, k \neq \ell} X_k X_\ell + \sum_{i,k,\ell \in \text{in}_G(j), i \neq k, k \neq \ell, \ell \neq i} X_i X_k X_\ell.$$

To examine the performance of PLACID with different values of γ, we set $\text{iv}_G(j) = \{j, j+p\}$ for all Y_j with descendants and vary γ among 1, 2, and 3. These values correspond to the cases where γ is less than, equal to, and larger than the number of valid IVs for certain primary variables. We set the sample size $n = 2000$ and the dimensions $(p, q, r) = (10, 30, 5)$.

The causal discovery and parameter estimation results are presented in Tables S5.12 and S5.13, respectively. We also report the empirical measures of surrogate IV strength in Table S5.14. Tables S5.12 and S5.13 show that PLACID performs best when γ is set to 2. In contrast, its accuracy decreases with $\gamma = 1$, particularly in settings with discrete secondary variables. This occurs because a smaller γ leads to fewer surrogate IVs according to Definition 3, which may lack sufficient strength. Table S5.14 confirms that the IV strength at $\gamma = 1$ is notably lower. In the discrete case, the first-stage F-statistic is below 10, corresponding to the rule-of-thumb threshold for weak IVs (Staiger and Stock, 1997). On the other hand, setting $\gamma = 3$ violates Assumption 3 as it exceeds the number of valid IVs for some primary variables. Consequently, the accuracy of PLACID decreases

with $\gamma = 3$, especially for parameter estimation.

Table S5.12: Means and standard deviations (in parentheses) of different causal discovery performance metrics for PLACID with continuous and discrete secondary variables across varying values of γ.

Setting	Graph	γ	FDP	TPR	SHD	JI
Continuous	Random	1	0.03(0.10)	0.95(0.14)	0.34(1.01)	0.94(0.17)
		2	0.01(0.04)	0.97(0.10)	0.11(0.37)	0.97(0.10)
		3	0.05(0.15)	0.92(0.20)	0.55(1.45)	0.90(0.22)
	Hub	1	0.00(0.00)	0.99(0.09)	0.09(0.81)	0.99(0.09)
		2	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		3	0.00(0.00)	0.98(0.12)	0.22(1.06)	0.98(0.12)
Discrete	Random	1	0.00(0.00)	0.80(0.34)	0.38(0.62)	0.80(0.34)
		2	0.01(0.05)	0.96(0.14)	0.23(0.71)	0.95(0.15)
		3	0.01(0.05)	0.91(0.22)	0.33(0.70)	0.90(0.22)
	Hub	1	0.00(0.00)	0.86(0.19)	1.25(1.72)	0.86(0.19)
		2	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		3	0.00(0.00)	0.97(0.14)	0.25(1.27)	0.97(0.14)

Table S5.13: Means and standard deviations (in parentheses) of different estimation losses for PLACID with continuous and discrete secondary variables across varying values of γ.

Setting	Graph	γ	L_∞	L_1	L_2
Continuous	Random	1	0.25(0.36)	0.61(1.33)	0.34(0.58)
		2	0.22(0.33)	0.45(0.85)	0.29(0.48)
		3	0.36(0.84)	0.84(1.91)	0.47(1.00)
	Hub	1	0.24(0.19)	1.03(1.03)	0.40(0.36)
		2	0.17(0.07)	0.68(0.24)	0.27(0.10)
		3	0.27(0.73)	1.17(3.28)	0.45(1.16)
Discrete	Random	1	0.62(0.44)	1.16(1.16)	0.78(0.62)
		2	0.16(0.30)	0.27(0.63)	0.19(0.38)
		3	0.50(0.21)	1.06(0.74)	0.68(0.33)
	Hub	1	0.87(0.30)	4.07(2.43)	1.61(0.80)
		2	0.09(0.03)	0.36(0.12)	0.14(0.04)
		3	0.48(0.03)	3.88(0.18)	1.30(0.06)

Table S5.14: Means of the average first-stage F-statistic as an empirical measure of surrogate IV strength for PLACID across varying values of γ.

Setting	Graph	γ	F-statistic
Continuous	Random	1	65.06
		2	273.73
		3	169.27
	Hub	1	29.95
		2	106.23
		3	64.43
Discrete	Random	1	9.84
		2	141.11
		3	191.26
	Hub	1	5.58
		2	57.09
		3	99.51

S5.5 Varying numbers of valid IVs

In this subsection, we conduct additional simulations to examine how PLACID performs with varying numbers of valid IVs for the primary variables.

We consider the same data-generating process as in Section S5.4 with $n = 2000$, $q = 3p$, $\mathrm{ca}_G(j) = \{j, j+p, j+2p\}$ for all $j = 1, \ldots, p$, and $(p, q, r) = (10, 30, 5)$. We set the valid IV set $\mathrm{iv}_G(j)$ to $\{j\}$, $\{j, j+p\}$, or $\{j, j+p, j+2p\}$ for all Y_j with descendants, corresponding to 1, 2, or 3 valid IVs. Following the setup in Section S5.4, for any X_ℓ that is a candidate IV of Y_j but not a valid IV of Y_j, we let X_ℓ have causal effects on all descendants of Y_j. To meet Assumption 3, we set γ to 1, 2, or 3 accordingly, so that $\gamma = \min_{1 \le j \le p} |\mathrm{iv}_G(j)|$. We then compare PLACID with the competing methods across these settings.

The causal discovery results for the continuous and discrete cases are presented in Tables S5.15 and S5.16, respectively, with parameter estimation results in Table S5.17. We also report the empirical measures of surrogate IV strength in Table S5.18. Tables S5.15–S5.17 show that the performance of PLACID improves as the number of valid IVs increases. On the other hand, its performance declines notably in settings where only one valid IV is available and the secondary variables are discrete, as shown in Tables S5.16 and S5.17. This is attributable to the much weaker surrogate

IVs in these settings, as reflected in Table S5.18. In particular, for the hub graph with discrete secondary variables, the average first-stage F-statistic falls below 10, meeting the conventional threshold for weak IVs (Staiger and Stock, 1997). Nevertheless, PLACID remains robust, controlling the FDP below the nominal level $q^* = 0.05$ and outperforming the competing methods across all scenarios.

Table S5.15: Means and standard deviations (in parentheses) of different causal discovery performance metrics for four methods with continuous secondary variables across varying numbers of valid IVs.

Graph	# valid IVs	Method	FDP	TPR	SHD	JI
Random	1	PLACID	0.03(0.10)	0.91(0.18)	0.38(0.81)	0.90(0.20)
		GrIVET	0.34(0.42)	0.52(0.41)	2.05(1.84)	0.43(0.38)
		RFCI	0.01(0.11)	0.32(0.37)	1.65(1.21)	0.32(0.37)
		LRpS-GES	0.20(0.33)	1.00(0.00)	0.46(0.66)	0.80(0.33)
	2	PLACID	0.01(0.04)	0.97(0.10)	0.11(0.37)	0.97(0.10)
		GrIVET	0.25(0.37)	0.50(0.39)	1.86(1.65)	0.43(0.38)
		RFCI	0.01(0.11)	0.46(0.43)	1.48(1.47)	0.46(0.43)
		LRpS-GES	0.13(0.27)	1.00(0.00)	0.36(0.67)	0.87(0.27)
	3	PLACID	0.01(0.06)	1.00(0.01)	0.08(0.27)	0.98(0.06)
		GrIVET	0.26(0.36)	0.54(0.39)	1.86(1.95)	0.45(0.38)
		RFCI	0.00(0.00)	0.69(0.37)	0.81(1.02)	0.69(0.37)
		LRpS-GES	0.15(0.34)	0.99(0.04)	0.33(0.93)	0.85(0.34)
Hub	1	PLACID	0.00(0.00)	0.98(0.08)	0.22(0.72)	0.98(0.08)
		GrIVET	0.51(0.40)	0.43(0.39)	9.97(6.56)	0.32(0.34)
		RFCI	0.41(0.37)	0.42(0.35)	7.01(3.84)	0.37(0.32)
		LRpS-GES	0.18(0.12)	0.85(0.13)	3.07(2.19)	0.73(0.18)
	2	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.32(0.39)	0.35(0.36)	8.38(5.36)	0.31(0.33)
		RFCI	0.17(0.21)	0.69(0.32)	4.15(3.45)	0.62(0.31)
		LRpS-GES	0.14(0.12)	0.80(0.13)	3.05(1.96)	0.72(0.17)
	3	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.23(0.35)	0.41(0.39)	6.71(4.67)	0.38(0.37)
		RFCI	0.07(0.23)	0.87(0.32)	1.53(3.53)	0.86(0.32)
		LRpS-GES	0.04(0.12)	0.81(0.31)	1.94(3.05)	0.80(0.32)

Table S5.16: Means and standard deviations (in parentheses) of different causal discovery performance metrics for four methods with discrete secondary variables across varying numbers of valid IVs.

Graph	# valid IVs	Method	FDP	TPR	SHD	JI
Random	1	PLACID	0.00(0.00)	0.74(0.39)	0.46(0.67)	0.74(0.39)
		GrIVET	0.46(0.43)	0.33(0.39)	2.74(2.25)	0.29(0.37)
		RFCI	0.00(0.00)	0.62(0.38)	1.08(1.25)	0.62(0.38)
		LRpS-GES	0.39(0.30)	0.96(0.13)	1.60(1.39)	0.61(0.31)
	2	PLACID	0.01(0.05)	0.96(0.14)	0.23(0.71)	0.95(0.15)
		GrIVET	0.20(0.39)	0.26(0.40)	2.15(1.68)	0.26(0.40)
		RFCI	0.00(0.00)	0.82(0.30)	0.52(0.94)	0.82(0.30)
		LRpS-GES	0.38(0.32)	0.99(0.08)	1.30(1.21)	0.62(0.32)
	3	PLACID	0.00(0.03)	1.00(0.00)	0.02(0.14)	1.00(0.03)
		GrIVET	0.07(0.23)	0.43(0.36)	1.60(1.46)	0.42(0.36)
		RFCI	0.00(0.00)	0.95(0.18)	0.14(0.38)	0.95(0.18)
		LRpS-GES	0.37(0.26)	0.98(0.08)	1.46(1.27)	0.62(0.27)
Hub	1	PLACID	0.00(0.00)	0.81(0.29)	1.67(2.62)	0.81(0.29)
		GrIVET	0.01(0.09)	0.26(0.42)	6.72(3.81)	0.26(0.42)
		RFCI	0.00(0.00)	0.05(0.06)	8.57(0.57)	0.05(0.06)
		LRpS-GES	0.19(0.26)	0.31(0.16)	6.99(1.88)	0.29(0.15)
	2	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.58(0.44)	0.12(0.18)	11.14(4.31)	0.11(0.16)
		RFCI	0.00(0.00)	0.79(0.15)	1.85(1.36)	0.79(0.15)
		LRpS-GES	0.15(0.19)	0.50(0.15)	5.33(2.03)	0.47(0.16)
	3	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.23(0.34)	0.25(0.22)	7.61(2.50)	0.23(0.21)
		RFCI	0.00(0.00)	0.81(0.38)	1.72(3.46)	0.81(0.38)
		LRpS-GES	0.05(0.14)	0.76(0.33)	2.35(3.31)	0.76(0.34)

Table S5.17: Means and standard deviations (in parentheses) of different estimation losses for two methods with continuous and discrete secondary variables across varying numbers of valid IVs.

Setting	Graph	# valid IVs	Method	L_∞	L_1	L_2
Continuous	Random	1	PLACID	0.46(0.42)	1.06(1.39)	0.61(0.65)
			GrIVET	0.87(0.49)	1.89(1.42)	1.15(0.69)
		2	PLACID	0.22(0.33)	0.45(0.85)	0.29(0.48)
			GrIVET	0.83(0.43)	1.62(1.33)	1.05(0.63)
		3	PLACID	0.06(0.10)	0.11(0.22)	0.07(0.13)
			GrIVET	0.73(0.46)	1.44(1.35)	0.92(0.65)
	Hub	1	PLACID	0.42(0.22)	1.62(0.89)	0.66(0.34)
			GrIVET	1.21(0.28)	11.13(5.60)	3.13(1.04)
		2	PLACID	0.17(0.07)	0.68(0.24)	0.27(0.10)
			GrIVET	1.08(0.28)	8.52(4.35)	2.65(0.97)
		3	PLACID	0.08(0.03)	0.30(0.09)	0.13(0.04)
			GrIVET	1.03(0.29)	6.77(4.00)	2.29(1.03)
Discrete	Random	1	PLACID	0.83(0.39)	1.68(1.22)	1.10(0.61)
			GrIVET	1.81(1.66)	3.70(3.30)	2.25(1.81)
		2	PLACID	0.16(0.30)	0.27(0.63)	0.19(0.38)
			GrIVET	0.98(0.47)	2.28(1.67)	1.38(0.73)
		3	PLACID	0.02(0.01)	0.03(0.02)	0.02(0.01)
			GrIVET	0.89(0.39)	1.96(1.81)	1.20(0.67)
	Hub	1	PLACID	0.61(0.40)	3.04(2.62)	1.18(0.92)
			GrIVET	1.88(1.13)	13.67(8.42)	4.64(2.81)
		2	PLACID	0.09(0.03)	0.36(0.12)	0.14(0.04)
			GrIVET	1.17(0.10)	10.82(2.50)	3.19(0.37)
		3	PLACID	0.02(0.01)	0.09(0.02)	0.04(0.01)
			GrIVET	1.15(0.43)	7.93(2.34)	2.71(0.67)

Table S5.18: Means of the average first-stage F-statistic as an empirical measure of surrogate IV strength for PLACID across varying numbers of valid IVs.

Setting	Graph	# valid IVs	F-statistic
Continuous	Random	1	61.30
		2	273.73
		3	282.70
	Hub	1	24.02
		2	106.23
		3	93.89
Discrete	Random	1	12.03
		2	141.11
		3	348.12
	Hub	1	6.64
		2	57.09
		3	193.45

S6 Additional analysis for the ADNI study

In this section, we provide additional analysis for the ADNI study. Specifically, Figure S6.1 visualizes potential nonlinear relationships between primary and secondary variables, while Figure S6.2 examines the independence among secondary variables.

To show the potential nonlinear relationships, we present the partial residual plots for several primary and secondary variables in both AD-MCI and CN groups. Partial residual plot is a frequently useful graphical diagnostic for nonlinearity among variables (Cook, 1993). A greater deviation between the solid and dashed lines in Figures S6.1 suggests a nonlinear relationship. It can be observed that some primary and secondary variables exhibit highly nonlinear relationships.

We further provide DC heatmaps for secondary variables in both AD-MCI and CN groups as shown in Figure S6.2. Since low DC values in Figure S6.2 indicate weak dependence among secondary variables, we empirically conclude that Assumption 1 approximately holds. Additionally, the heatmaps from two groups show similar patterns, suggesting similar SNP dependencies across the dataset.

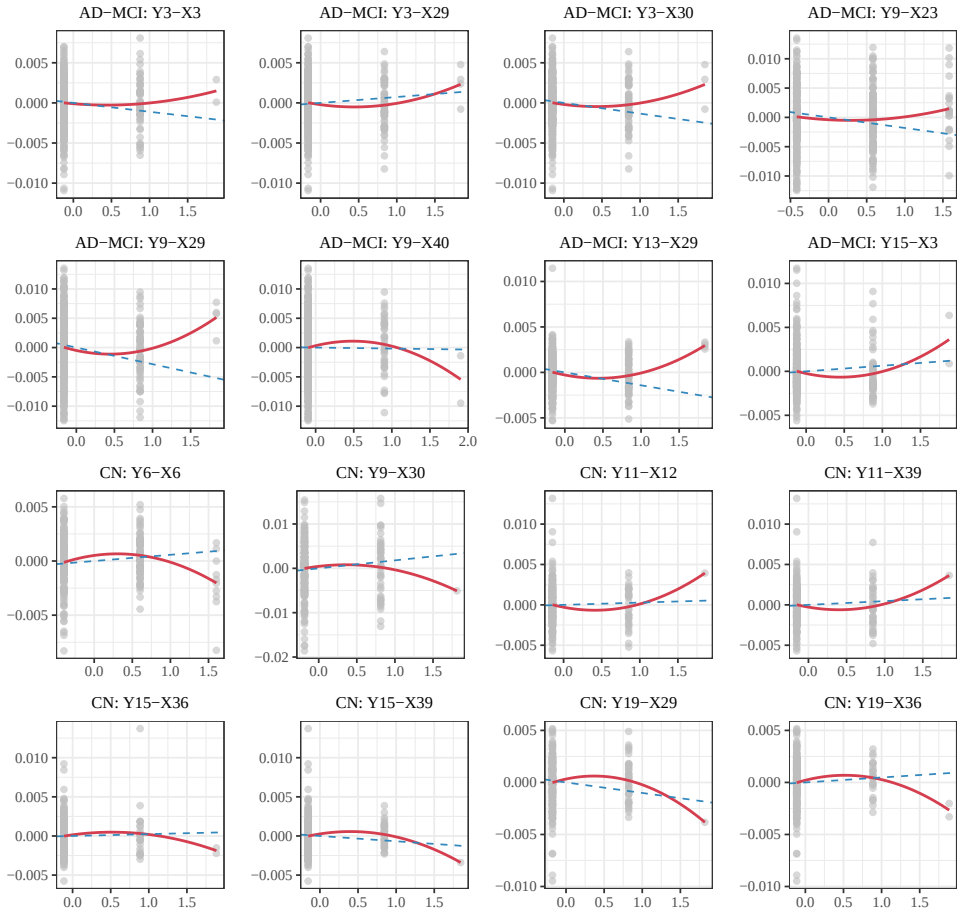
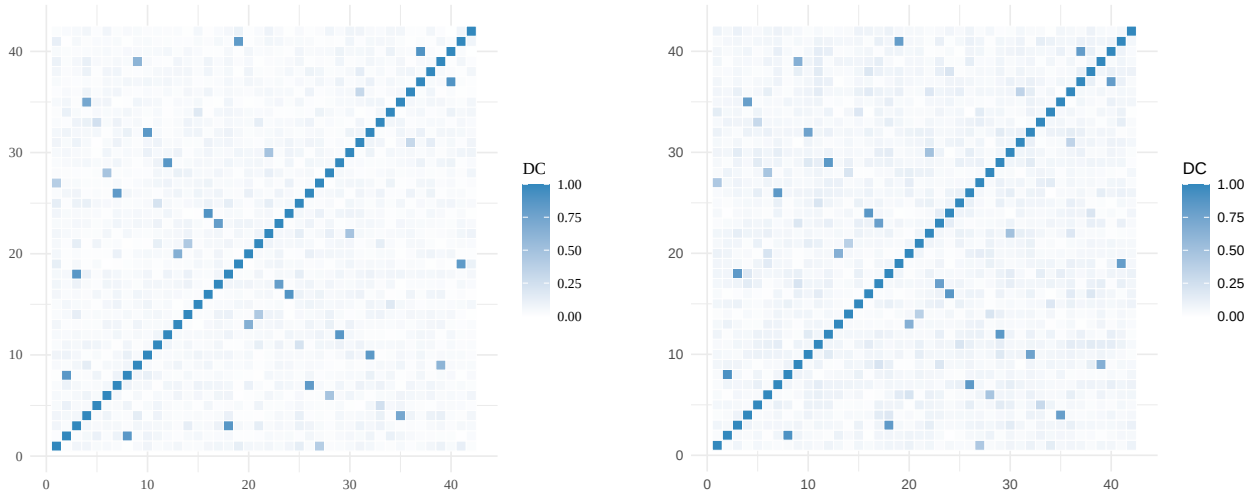


Figure S6.1: Partial residual plots for AD-MCI and CN groups.



(a) AD-MCI

(b) CN

Figure S6.2: Heatmaps of DC for (a) AD-MCI and (b) CN groups.

Bibliography

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165–1188.

Chen, L., C. Li, X. Shen, and W. Pan (2024). Discovery and inference of a causal network with hidden confounding. *Journal of the American Statistical Association 119*(548), 2572–2584.

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics 35*(4), 351–362.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.