Statistica Sinica Preprint No: SS-2025-0331	
Title	Semiparametric Causal Discovery and Inference with Invalid Instruments
Manuscript ID	SS-2025-0331
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0331
Complete List of Authors	Jing Zou, Wei Li and Wei Lin
Corresponding Authors	Wei Li
E-mails	weilistat@ruc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

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*

Abstract: Learning causal relationships among a set of variables, as encoded by a directed acyclic graph, from observational data is complicated by the presence of unobserved confounders. Instrumental variables (IVs) are a popular remedy for this issue, but most existing methods either assume the validity of all IVs or postulate a specific form of relationship, such as a linear model, between the primary variables and the IVs. To overcome these limitations, we introduce a partially linear model for causal discovery and inference that accommodates potentially invalid IVs and allows for general dependence of the primary variables on the IVs. We establish identification under this semiparametric model by constructing surrogate valid IVs, and develop a finite-sample procedure for estimating the causal structures and effects. Theoretically, we show that our procedure consistently learns the causal structures, yields asymptotically normal estimates, and effectively controls the false discovery rate in edge recovery. Simulation studies demonstrate the superiority of our method over existing competitors, and an application to inferring gene regulatory networks in Alzheimer's disease illustrates its usefulness.

Key words and phrases: causal inference; directed acyclic graph; identification; instrumental variable; unobserved confounding.

[*]Corresponding authors

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

1. Introduction

Inferring causal relationships among a set of variables, as represented by a directed acyclic graph (DAG), is a fundamental problem in statistics and finds applications in various fields such as systems biology (Triantafillou et al., 2017) and medical imaging (Castro et al., 2020). As a result, learning causal relationships from observational data, known as causal discovery, has gained popularity and become an active area of research; see Heinze-Deml et al. (2018) for comprehensive reviews. The potential existence of unobserved confounders, however, may cause violations of the Markov property in the DAG and lead to biased estimates of causal effects (Pearl, 2009), posing challenges to robust causal discovery. The existing literature offers some approaches to tackling this challenge. One way is to generate less informative discoveries, such as constructing a partial ancestor graph instead of a DAG (Colombo et al., 2012). Another way is to develop effective algorithms for deconfounding, often under additional assumptions about the confounding mechanism. For example, Li et al. (2024) proposed a deconfounded estimation procedure for nonlinear causal discovery under a sublinear growth condition that separates linear confounding effects from nonlinear causal relationships. To avoid such restrictions while obtaining an accurate estimate of the DAG, here we resort to the use of instrumental variables (IVs), which provides a convenient and powerful method for resolving the issue of unobserved confounding.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

In the classical context of inferring a treatment–outcome relationship, there has been a rich body of work on estimating causal effects using valid IVs (e.g., Angrist et al., 1996; Newey and Powell, 2003). However, owing to the untestable independence and exclusion restriction assumptions, candidate IVs may be invalid, rendering the estimates biased or inconsistent. In the presence of invalid IVs, Bowden et al. (2015) and Kolesár et al. (2015) identified causal effects by assuming that the direct effects of the IVs on the outcome are asymptotically orthogonal to their effects on the treatment. Alternatively, some studies made assumptions regarding the proportion of valid IVs among the candidate set. In particular, Kang et al. (2016) and Windmeijer et al. (2019) developed Lasso-based methods for selecting valid IVs and estimating causal effects under the majority rule, which requires more than half of the candidate IVs to be valid. Guo et al. (2018) proposed two-stage hard thresholding with voting for constructing confidence intervals under the more general plurality rule. Recently, Sun et al. (2023) introduced a new class of G-estimators for a semiparametric structural equation model (SEM), allowing for a flexible number of valid IVs and bypassing the IV selection step.

Despite extensive developments on IV methods in causal inference, their application to causal discovery remains underexplored. Notably, with knowing a priori a unique valid IV for each primary variable, Oates et al. (2016) formalized

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

the notion of conditional DAGs and developed a score-based estimation method via integer linear programming. Among the few attempts to learn a causal graph with invalid IVs (Chen et al., 2023; Li et al., 2023; Chen et al., 2024; Li et al., 2024), Chen et al. (2023) proposed a stepwise IV selection procedure followed by two-stage least squares and Wald tests for inference, while Li et al. (2023) and Chen et al. (2024) developed peeling algorithms to estimate ancestral relation graphs and candidate IV sets, along with likelihood-based edge inference. Still, all these methods rely on the assumption of a linear SEM and do not account for candidate IVs whose effects on the primary variables may be nonlinear. As widely recognized in the literature, failing to exploit such nonlinearities may result in weak IVs, deteriorate the estimation performance, or distort the causal interpretation (Newey, 1990; Sun et al., 2023).

In this paper, we consider causal discovery and inference in the presence of unobserved confounders using potentially invalid IVs. Unlike existing studies that postulate linear relationships between the primary variables and the IVs, we adopt a partially linear SEM that leaves the functional form of these relationships unspecified. Within this semiparametric framework, we establish identification of causal structures and effects under relatively mild assumptions. Specifically, we first identify the ancestral relationships and candidate IV sets by extending the peeling algorithm of Chen et al. (2024) to our more general setting. Building on

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

these results, we then construct surrogate valid IVs and derive moment conditions to identify the causal effects recursively, which generalizes the identification strategy of Sun et al. (2023) to causal graphs. We further develop a finite-sample procedure for causal discovery and inference by using distance-correlation-based independence tests and the generalized method of moments. We call our method the Partially Linear Approach to Causal Instrument-based Discovery (PLACID). Theoretically, we show that PLACID consistently learns the causal structures, yields asymptotically normal estimates, and effectively controls the false discovery rate in edge recovery.

The remainder of this paper is organized as follows. Section 2 introduces our causal graph terminology and the partially linear SEM. Section 3 establishes the identification results for the causal graph and causal effects. Section 4 presents the PLACID methodology along with its theoretical guarantees. Sections 5 and 6 illustrate the numerical performance of our method through simulation studies and an application to an Alzheimer's disease dataset, respectively. Section 7 concludes the paper with some discussion.

We introduce some notation that will be used throughout the paper. For a K-dimensional random variable $\mathbf{Z}$ and an index set $\alpha \subseteq \{1, \ldots, K\}$, define $\mathbf{Z}_\alpha = (Z_s : s \in \alpha)$ and $\mathbf{Z}_{-\alpha} = (Z_s : s \notin \alpha)$. Let $\mathbb{Z}$ denote the data realizations of $\mathbf{Z}$. Let $\mathcal{H}(\mathbf{Z})$ denote the Hilbert space of one-dimensional functions of $\mathbf{Z}$ with

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

mean zero and finite variance, equipped with the covariance inner product. Let $|J|$ denote the cardinality of a set J. For a matrix $\mathbf{A} = (A_{ij})$, let $\mathbf{A}_{i,\cdot}$ denote its ith row and $\mathbf{A}_{\cdot,j}$ its jth column. For a vector $\mathbf{v}$, let $\mathbf{v}^T$ denote its transpose and $\|\mathbf{v}\|_0$ its L_0-norm. Let $\mathbf{1}(\cdot)$ denote the indicator function and $\mathbf{I}_p$ the $p \times p$ identity matrix. Finally, let $\hat{E}_n$ denote the empirical mean operator with respect to a sample of size n.

2. Causal graphical model

Consider a causal graph G with p endogenous primary variables $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$ and q exogenous secondary variables $\mathbf{X} = (X_1, \ldots, X_q)^T$, both having finite variance. Specifically, we denote

$$G = (\mathbf{X}, \mathbf{Y}; \mathcal{E}, \mathcal{I}), \tag{2.1}$$

where $\mathcal{E} = \{(i, j) : Y_i \to Y_j\}$ is the set of directed edges among $\mathbf{Y}$, and $\mathcal{I} = \{(\ell, j) : X_\ell \to Y_j\}$ is the set of directed edges from $\mathbf{X}$ to $\mathbf{Y}$. Note that there is no directed edge from $\mathbf{Y}$ to $\mathbf{X}$, and thus $\mathbf{X}$ can be viewed as external interventions. Based on G, we adopt the following terminology: (i) the parent set of Y_j, $\mathrm{pa}_G(j) = \{k : Y_k \to Y_j\}$; (ii) if there exists a directed path from Y_k to Y_j, then Y_j is a descendant of Y_k, Y_k is an ancestor of Y_j, and the ancestor set of Y_j is $\mathrm{an}_G(j) = \{k : Y_k \to \cdots \to Y_j\}$; (iii) the intervention set of Y_j, $\mathrm{in}_G(j) = \{\ell : X_\ell \to Y_j\}$;

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

(iv) the mediator set of Y_k and Y_j, $\text{me}_G(k,j) = \{i : Y_k \to \cdots \to Y_i \to \cdots \to Y_j\}$; (v) the non-mediator set of Y_k and Y_j, $\text{nm}_G(k,j) = \text{an}_G(j) \setminus [\text{me}_G(k,j) \cup \{k\}]$; (vi) Y_k is an unmediated parent of Y_j if $(k,j) \in \mathcal{E}$ and $\text{me}_G(k,j) = \emptyset$; (vii) the leaf nodes of G, $\text{leaf}(G) = \{j : Y_j \text{ has no descendant in } G\}$; (viii) the ancestral relation graph (ARG), $G^+ := (\mathbf{X}, \mathbf{Y}; \mathcal{E}^+, \mathcal{I}^+)$, where $\mathcal{E}^+ = \{(k,j) : k \in \text{an}_G(j)\}$, $\mathcal{I}^+ = \big\{(\ell, j) : \ell \in \bigcup_{k \in \text{an}_G(j) \cup \{j\}} \text{in}_G(k)\big\}$. 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 exists a directed path from X_ℓ to Y_j in G, then $(\ell, j) \in \mathcal{I}^+$. To further illustrate these definitions, we consider a simple case in Example S1.1 of the Supplementary Material.

Following the idea of conditional DAGs (Oates et al., 2016), we wish to use $\mathbf{X}$ as candidate IVs to infer the causal relationships and effects among $\mathbf{Y}$. The variables $\mathbf{X}$ are exogenous, implying that these variables satisfy the independence assumption of valid IVs. There are directed edges from $\mathbf{X}$ to $\mathbf{Y}$, but none in the opposite direction, providing the basis for the relevance assumption. Therefore, it is possible to use $\mathbf{X}$ as IVs. We further introduce our definitions of valid and candidate IVs based on causal graphs.

Definition 1 (Valid IV)**.** A secondary variable X_ℓ is said to be a valid IV for Y_j in the causal graph G, if it intervenes on Y_j, i.e., $(\ell, j) \in \mathcal{I}$ and does not intervene on any other primary variable Y_i, i.e., $(\ell, i) \notin \mathcal{I}$ for all $i \neq j$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

Definition 2 (Candidate IV)**.** A secondary variable X_ℓ is said to be a candidate IV for Y_j in the causal graph G, if it intervenes on Y_j, i.e., $(\ell, j) \in \mathcal{I}$ and does not intervene on any non-descendant of Y_j.

Accordingly, denote the set of valid IVs for Y_j in G by $\mathrm{iv}_G(j) = \{\ell : X_\ell \to Y_j, X_\ell \nrightarrow Y_i, i \neq j\}$ and the set of candidate IVs for Y_j in G by $\mathrm{ca}_G(j) = \{\ell : X_\ell \to Y_j, X_\ell \to Y_k \text{ only if } j \in \mathrm{an}_G(k)\}$. It is obvious that $\mathrm{iv}_G(j) \subseteq \mathrm{ca}_G(j)$. However, a candidate IV may not be valid, because it may intervene on a descendant of Y_j, which contradicts Definition 1.

Since the variables $\mathbf{Y}$ are of primary interest, it is often reasonable to assume simple relationships among $\mathbf{Y}$ while leaving unrestricted the functional forms of interactions between $\mathbf{X}$ and $\mathbf{Y}$, as shown useful in semiparametric modeling in econometrics (Engle et al., 1986) and environmental science (Dominici et al., 2004). Among such semiparametric models, of particular importance is the partially linear model (PLM) (Robinson, 1988). For the causal graph G in (2.1), we consider the partially linear SEM

$$Y_j = \sum_{i=1}^{p} \beta_{ij}^* Y_i + g_j(\mathbf{X}_{\mathrm{in}_G(j)}) + \varepsilon_j, \quad E(\varepsilon_j) = 0, \quad \mathbf{X} \perp\!\!\!\perp \varepsilon_j, \quad j = 1, \ldots, p. \quad (2.2)$$

Here β_{ij}^* represents the direct causal effect of Y_i on Y_j, and $\beta_{ij}^* \neq 0$ implies that Y_i is a cause of Y_j, i.e., $i \in \mathrm{pa}_G(j)$. The function $g_j(\cdot)$ captures the causal effect of $\mathbf{X}_{\mathrm{in}_G(j)}$ on Y_j, the form of which is unknown and not restricted to linear

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

functions. Our interest lies in estimating the causal structures and effects among $\mathbf{Y}$, as characterized by the edge set $\mathcal{E}$ and the coefficient matrix $\mathbf{B}^* = (\beta^*_{ij})_{p\times p}$.

Compared to the standard linear model, the inclusion of a nonparametric term in (2.2) enhances robustness against model misspecification (Florens et al., 2012). Previous research has extensively examined the use of PLMs in causal inference, including estimation with missing data (Liang et al., 2004) and identifiability of partially linear SEMs (Rothenhäusler et al., 2018). However, since potential unobserved confounders have been absorbed in $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$, $\mathrm{Cov}(\mathbf{Y}_{-\{j\}}, \varepsilon_j)$ may not be $\mathbf{0}$ in model (2.2). As a result, existing estimation methods for PLMs are unsuitable in our context, even when the causal relationships are known. Assuming that $g_j(\cdot)$ takes a linear form, Chen et al. (2024) proposed an insightful peeling algorithm to infer the causal relationships and effects among the primary variables $\mathbf{Y}$ using IVs. Their method, while effective in linear settings, may not be applicable to the more general semiparametric model (2.2), because an IV deemed valid under Definition 1 might not satisfy the criteria for a valid IV in Chen et al. (2024). Without valid IVs, the causal parameters in their model will become unidentifiable. Specifically, Chen et al. (2024) defined X_ℓ as a valid IV for Y_k if $W_{\ell k} \neq 0$ and $W_{\ell k'} = 0$ for all $k' \neq k$, where $\mathbf{W} = (W_{\ell k})_{q\times p} = \mathbf{V}(\mathbf{I}_p - \mathbf{B}^*)$ and $\mathbf{V}$ is the linear regression coefficient of $\mathbf{Y}$ on $\mathbf{X}$, i.e., $\mathbf{V} = \{\mathrm{Var}(\mathbf{X})\}^{-1}\,\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$. The following example clearly demonstrates this point.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

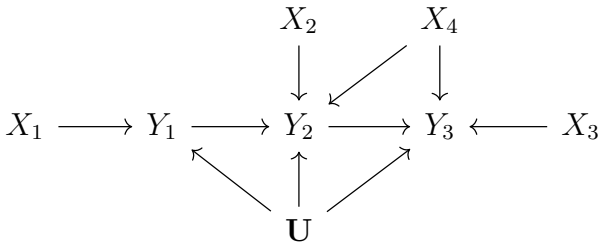


Figure 1: An example of the causal graph G.

Example 1. Consider the causal graph G shown in Figure 1, where $\mathbf{U}$ denotes unobserved confounders, $Y_1 = X_1^2 + \varepsilon_1$, Y_j follows model (2.2) with unspecified $g_j(\cdot)$ for $j = 2, 3$, and $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_4)$. It is easy to verify that $\text{cov}(X_1, \mathbf{Y}) = \mathbf{0}$, and hence $\mathbf{V}_{1,\cdot} = \mathbf{0}$ and $\mathbf{W}_{1,\cdot} = \mathbf{0}$. This shows that, although X_1 is indeed a valid IV for Y_1 in G, it is not a valid IV for any primary variable in $\mathbf{Y}$ under Chen et al. (2024). Consequently, Y_1 is considered to lack valid IVs, and thus the approach in Chen et al. (2024) fails to identify the causal parameters in the model for Y_1.

In light of Example 1, the method of Chen et al. (2024) fails to identify the causal parameters in model (2.2). In contrast, our approach introduced in the next section still ensures identification of these parameters. Furthermore, when $g_j(\cdot)$ is unknown and possibly highly nonlinear, the use of linear methods can introduce substantial estimation bias. This issue may be particularly critical for the purpose of causal discovery since it can easily lead to incorrect determination of causal directions among the primary variables. To address this problem, we next consider the discovery and estimation of DAGs using invalid IVs under model

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

(2.2), which has not been discussed in the existing literature.

3. Identification of the causal graph and effects

In this section, we show how to identify the causal structures and effects among the primary variables $\mathbf{Y}$ using the secondary variables $\mathbf{X}$. We first make the following assumptions.

Assumption 1. The secondary variables are independent of each other, i.e., $X_i \perp\!\!\!\perp (X_j)_{j\neq i}$ for all $i = 1, \ldots, q$.

Assumption 1 is reasonable and common in Mendelian randomization studies, where genetic variants from different gene regions are often used as IVs. To meet this assumption, one can employ well-established tools for linkage disequilibrium clumping such as PLINK (Purcell et al., 2007) to select independent genetic variants. This practice is prevalent and generally accepted within the field (e.g., Zhao et al., 2020; Ye et al., 2021). The independence assumption has also been made in existing methods for causal discovery. For instance, Neto et al. (2010) and Ongen et al. (2017) treated expression quantitative trait loci (eQTLs) as secondary variables and exploited their independence when evaluating genetic causality.

Assumption 2. Whenever X_ℓ intervenes on an unmediated parent of Y_j, $X_\ell \not\perp\!\!\!\perp Y_j$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

Assumption 2 is the faithfulness assumption in causal discovery (Spirtes et al., 2001). It requires that when X_ℓ intervenes on both Y_j and its unmediated parent, the dependency between X_ℓ and Y_j should not be canceled out.

Assumption 3. For each primary variable Y_j, there are at least $\gamma \geq 1$ valid IVs, i.e., $|\operatorname{iv}_G(j)| \geq \gamma$ for all $j = 1, \ldots, p$.

When there are only two primary variables with known causal direction, Assumption 3 is the same as the one imposed in Sun et al. (2023). The value of γ can be specified based on prior knowledge. In general, it is possible to set $\gamma = 1$ by assuming only the existence of a valid IV for each Y_j, as done by Li et al. (2023) and Zilinskas et al. (2024). However, these studies did not account for unobserved confounders. In the presence of unobserved confounding, Chen et al. (2024) assumed the majority rule as in Kang et al. (2016), which is stronger than Assumption 3 since it requires at least half of the candidate IVs to be valid for each primary variable.

For causal inference between an exposure and an outcome of interest under a semiparametric model similar to (2.2), Sun et al. (2023) introduced a system of moment conditions for identification in a union of causal models where at least γ of the candidate IVs are valid but their identities are unknown. They then proposed a class of G-estimators (Robins et al., 1992) and developed semiparametric efficiency theory. To extend this idea to our causal graph setting, we first give the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

following definition. Let $d(\mathbf{X}_{\mathrm{ca}_G(j)})$ denote a generic one-dimensional function of $\mathbf{X}_{\mathrm{ca}_G(j)}$ with mean zero and finite variance, i.e., $d(\mathbf{X}_{\mathrm{ca}_G(j)}) \in \mathcal{H}(\mathbf{X}_{\mathrm{ca}_G(j)})$.

Definition 3. For each primary variable Y_j and an index set $\alpha_j \subseteq \mathrm{ca}_G(j)$ with $|\alpha_j| \geq \gamma$, define the subspace

$$\mathcal{D}(\alpha_j) = \big\{d(\mathbf{X}_{\mathrm{ca}_G(j)}) \in \mathcal{H}(\mathbf{X}_{\mathrm{ca}_G(j)}) : E\{d(\mathbf{X}_{\mathrm{ca}_G(j)}) \mid \mathbf{X}_{\mathrm{ca}_G(j)\setminus\alpha_j}\} = 0\big\}.$$

Further, define the intersection of all possible $\mathcal{D}(\alpha_j)$ by

$$\mathcal{Z}_\gamma(j) = \bigcap_{\alpha_j \subseteq \mathrm{ca}_G(j), |\alpha_j| \geq \gamma} \mathcal{D}(\alpha_j).$$

In particular, since $\Pi_{\ell \in \mathrm{ca}_G(j)}\{X_\ell - E(X_\ell)\}$ belongs to $\mathcal{Z}_\gamma(j)$ under Assumptions 1 and 3, $\mathcal{Z}_\gamma(j)$ is non-empty. For discrete $\mathbf{X}_{\mathrm{ca}_G(j)}$, the space $\mathcal{H}(\mathbf{X}_{\mathrm{ca}_G(j)})$ can be spanned by a finite number of orthogonal functions. Since $\mathcal{Z}_\gamma(j) \subseteq \mathcal{H}(\mathbf{X}_{\mathrm{ca}_G(j)})$, we can stack all basis functions of $\mathcal{Z}_\gamma(j)$ into a vector denoted by $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(j)})$. For example, consider the case where there is only one binary variable X_ℓ that is both the candidate and valid IV for Y_j, i.e., suppose that $\gamma = 1$, $\mathrm{ca}_G(j) = \mathrm{iv}_G(j) = \{\ell\}$, and $X_\ell \in \{0, 1\}$. Since X_ℓ is binary, $\mathcal{H}(X_\ell) = \mathrm{span}(\{X_\ell - E(X_\ell)\})$. Thus, by Definition 3, $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(j)})$ is exactly the centered valid instrument $X_\ell - E(X_\ell)$ for Y_j. Cases with multiple IVs are illustrated in Example S1.2 of the Supplementary Material. For continuous $\mathbf{X}_{\mathrm{ca}_G(j)}$, however, $\mathcal{Z}_\gamma(j)$ becomes an infinite-dimensional Hilbert space. This difficulty arises even in the simplest scenario with one candidate IV X_ℓ for Y_j, where $\mathcal{Z}_\gamma(j) = \big\{d(X_\ell) \in \mathcal{H}(X_\ell) : E\{d(X_\ell)\} = 0\big\}$ remains

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

infinite-dimensional. To address this issue, we follow the general strategy in Newey (1993) and Tchetgen Tchetgen et al. (2010) by selecting a basis set of functions $\{\phi_s(\mathbf{X}_{\mathrm{ca}_G(j)})\}_{s=1}^{\infty}$ that are dense in $\mathcal{Z}_\gamma(j)$, such as tensor products of trigonometric, wavelet, or polynomial bases. We can then construct $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(j)})$ from a finite subset of these basis functions.

Below we briefly explain how $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(j)})$ can serve as surrogate IVs for identification. Consider the simple case where $\mathrm{me}_G(k,j) = \emptyset$, meaning that the total causal effect of Y_k on Y_j is the direct effect β^*_{kj}. The causal relationships among the relevant variables for this case are illustrated in Figure 2, with dashed lines indicating that the relationships may exist. When Assumption 1 holds, it is easy to see that $\mathbf{Y}_{\mathrm{nm}_G(k,j)} \perp\!\!\!\perp \mathbf{X}_{\mathrm{ca}_G(k)}$ and $\mathbf{X}_{\mathrm{in}_G(j)} \perp\!\!\!\perp \mathbf{X}_{\mathrm{ca}_G(k)} \mid \mathbf{X}_{\mathrm{ca}_G(k)\setminus \mathrm{iv}_G(k)}$. Then, for any $d(\mathbf{X}_{\mathrm{ca}_G(k)}) \in \mathcal{D}\{\mathrm{iv}_G(k)\}$, we have

$$E\{d(\mathbf{X}_{\mathrm{ca}_G(k)})(Y_j - \beta^*_{kj} Y_k)\} = 0. \tag{3.1}$$

See (S3.1) in the Supplementary Material for the full derivation of (3.1). Notably, the nonparametric term $g_j(\mathbf{X}_{\mathrm{in}_G(j)})$ in the PLM model (2.2) is not contained in (3.1). The proposed approach differs from conventional PLM methods such as smoothing spline regression (Härdle et al., 2000), which approximate nonparametric terms via linear basis expansions. By leveraging the orthogonality between $g_j(\mathbf{X}_{\mathrm{in}_G(j)})$ and $\mathcal{D}\{\mathrm{iv}_G(k)\}$, the moment condition (3.1) is derived without ap-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

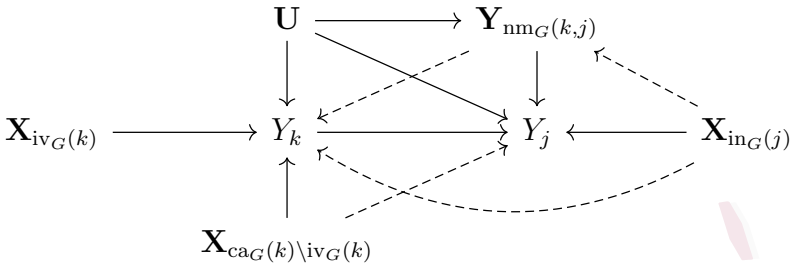


Figure 2: Causal relationships in the case where $\mathrm{me}_G(k,j) = \emptyset$.

proximating $g_j(\cdot)$, thereby preventing the approximation errors. Although $\mathrm{iv}_G(k)$ is unknown, there must exist a subset $\alpha_k \subseteq \mathrm{ca}_G(k)$ as described in Definition 3 such that $\alpha_k = \mathrm{iv}_G(k)$. By the definition of $\mathcal{Z}_\gamma(k)$, we have $\mathcal{Z}_\gamma(k) \subseteq \mathcal{D}\{\mathrm{iv}_G(k)\}$. Consequently, all random variables in $\mathcal{Z}_\gamma(k)$ must satisfy the moment condition (3.1), or equivalently $E\{\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})(Y_j - \beta_{kj}^* Y_k)\} = \mathbf{0}$ for the basis functions $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$ of $\mathcal{Z}_\gamma(k)$. In other words, we can construct surrogate IVs $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$ based on $\mathbf{X}_{\mathrm{ca}_G(k)}$ without the need to know the valid IVs.

The above discussion relies primarily on Assumption 1. Violations of this assumption may break the orthogonality between the nonparametric term $g_j(\mathbf{X}_{\mathrm{in}_G(j)})$ and the subspace $\mathcal{D}\{\mathrm{iv}_G(k)\}$, which can render the surrogate IVs invalid. A concrete example illustrating this scenario is provided in Example S1.3 of the Supplementary Material.

Assumption 4. For each primary variable Y_k with descendants,

$$\|E\{\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})Y_k\}\|_0 > 0.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

Since $E\{\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})\} = \mathbf{0}$ by Definition 3, Assumption 4 stems from the necessity for the surrogate IVs $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$ to satisfy the relevance assumption. Additionally, because $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$ characterizes the function space $\mathcal{Z}_\gamma(k)$, Assumption 4 entails the existence of a random variable in $\mathcal{Z}_\gamma(k)$ that is correlated with Y_k. This requirement is not stringent, in view of the fact that all the candidate IVs of Y_k are correlated with Y_k. Moreover, as γ increases, this assumption becomes milder. As an illustration, consider the case where all candidate IVs take values 0 or 1. If we believe that $m = |\,\mathrm{ca}_G(k)|$ candidate IVs of Y_k are all valid, then there are $2^m - 1$ basis functions in $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$, and Assumption 4 requires at least one of these basis functions to be correlated with Y_k. We are now ready to state our main identification result.

Theorem 1. *Suppose that Assumptions 1–4 hold. For the causal graph* $G = (\mathbf{X}, \mathbf{Y}; \mathcal{E}, \mathcal{I})$*, the edge set* $\mathcal{E}$ *and the causal parameters* $\{\beta^*_{ij}\}_{i\in \mathrm{pa}_G(j)}$ *in model* (2.2) *are identifiable.*

The proof of Theorem 1 exploits a two-stage identification strategy. We begin by identifying the ARG to roughly capture the causal directions among $\mathbf{Y}$ and obtain candidate IV sets. Specifically, Assumptions 1 and 2 imply that only valid IVs for leaf nodes depend exclusively on one primary variable while remaining independent of all others. This property enables the identification of leaf nodes and their IVs in G. Removing these nodes yields a subgraph whose leaf nodes

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

and the corresponding IVs remain identifiable via the same approach. Iteratively repeating this process enables the recovery of the entire ARG along with the candidate IV sets. Building on these results, we proceed to construct surrogate IVs to identify the causal effects and graph. Since the causal effect β_{kj} can be nonzero only when $(k, j) \in \mathcal{E}^+$, it suffices to identify $\boldsymbol{\beta}^* := (\beta^*_{kj})_{(k,j)\in\mathcal{E}^+}$. By extending the aforementioned surrogate IV framework to the entire graph, we establish that under Assumptions 1–4, $\boldsymbol{\beta}^*$ can be identified as the unique solution to

$$E\{\mathbf{M}(\boldsymbol{\beta}^*)\} = \mathbf{0}, \tag{3.2}$$

where $\mathbf{M}(\boldsymbol{\beta}^*)$ is the concatenation of all $M_{kj}(\boldsymbol{\beta}^*)$ for $(k, j) \in \mathcal{E}^+$ and

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

The detailed proof is given in Section S3.1 of the Supplementary Material.

4. Methodology and theory

In this section, we introduce PLACID, a finite-sample method for estimating the causal graph and causal effects in model (2.2). The method consists of an algorithm for estimating the ARG and candidate IV sets (Section 4.1) and a procedure for inferring the causal effects and directions (Section 4.2). Theoretical guarantees are provided in Section 4.3.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

4.1 Estimation of the ARG and candidate IV sets

To estimate the ARG, it is essential to establish the dependence between $\mathbf{X}$ and $\mathbf{Y}$ based on finite samples, which can be achieved using distance correlation (Székely et al., 2007). The distance correlation (DC) measures the dependence between two random vectors using the distance between their characteristic functions. Unlike Pearson's correlation coefficient, the DC is zero only if the random vectors are independent. Moreover, in the bivariate normal case, it is strictly increasing with the absolute value of Pearson's correlation coefficient. The notion of DC has been effectively used in feature screening (Li et al., 2012) and causal discovery from time series (Runge et al., 2019). Here we employ the empirical DC to test and measure the dependence of any pair X_i and Y_j. Calculation details and a full description of DC are provided in Section S2 of the Supplementary Material.

Algorithm 1 implements the identification strategy developed earlier in a finite-sample setting. Under Assumptions 1 and 2, each leaf node admits valid IVs that depend exclusively on that node and are independent of all others. This property allows us to iteratively identify and remove leaf nodes, along with recovering their relationships to the primary variables removed in earlier steps, thereby reconstructing the ARG. Algorithm 1 operationalizes this logic by using the empirical DC to estimate the dependence structure between $\mathbf{X}$ and $\mathbf{Y}$, as described by the iterative steps in lines 4–14. At each iteration, it selects a secondary vari-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

Algorithm 1 DC-based estimation of the ARG and candidate IV sets

Input: Data $(\mathbb{X}_{n\times q}, \mathbb{Y}_{n\times p})$, significance level $\alpha > 0$

Output: Estimates of $\mathcal{E}^+$, $\mathcal{I}^+$, and candidate IV sets

1: Compute the empirical DC matrix $\mathbf{C} = \{\mathcal{R}_n(X_i, Y_j)\}_{q\times p}$ via (S2.1) in the Supplementary Material

2: Compute the rejection matrix $\mathbf{R} = (R_{ij})_{q\times p}$ via (S2.2) in the Supplementary Material

3: Initialize $\hat{\mathcal{E}}^+ \leftarrow \emptyset$, $\hat{\mathcal{I}}^+ \leftarrow \{(i,j) : R_{ij} \neq 0\}$, $\mathbf{Y} \leftarrow \{1,\ldots,p\}$, $\mathbf{X} \leftarrow \{1,\ldots,q\}$, $\mathbf{Y}^- \leftarrow \mathbf{Y}$, $\mathbf{X}^- \leftarrow \mathbf{X}$, $\mathcal{E}^- \leftarrow \hat{\mathcal{E}}^+$, $\mathcal{I}^- \leftarrow \hat{\mathcal{I}}^+$, $\mathbf{R}^- \leftarrow \mathbf{R}$

4: **while** $\mathbf{Y}^- \neq \emptyset$ **do**

5: Initialize $\text{leaf}(G^-) \leftarrow \emptyset$, $\text{iv}_{G^-}(j) \leftarrow \emptyset$ for all $j \in \mathbf{Y}^-$

6: **for** $\ell \in \arg\min_{j:\|\mathbf{R}^-_{j,\cdot}\|_0>0} \|\mathbf{R}^-_{j,\cdot}\|_0$ **do**

7: $k \leftarrow \arg\max_{j:j\in\mathbf{Y}^-} C_{\ell j}$

8: $\text{leaf}(G^-) \leftarrow \text{leaf}(G^-) \cup \{k\}$

9: $\text{iv}_{G^-}(k) \leftarrow \text{iv}_{G^-}(k) \cup \{\ell\}$

10: **end for**

11: $\hat{\mathcal{E}}^+ \leftarrow \hat{\mathcal{E}}^+ \bigcup \{(k,j) : k \in \text{leaf}(G^-), j \in \mathbf{Y} \setminus \mathbf{Y}^-, R_{\ell j} \neq 0 \text{ for all } \ell \in \text{iv}_{G^-}(k)\}$

12: $\mathbf{Y}^- \leftarrow \mathbf{Y}^- \setminus \text{leaf}(G^-)$, $\mathbf{X}^- \leftarrow \mathbf{X}^- \setminus \bigcup_{k\in\text{leaf}(G^-)} \text{iv}_{G^-}(k)$

13: Update $\mathbf{R}^-$ by keeping the rows in $\mathbf{X}^-$ and columns in $\mathbf{Y}^-$

14: **end while**

15: $\hat{\mathcal{E}}^+ \leftarrow \{(k,j) : Y_k \to \cdots \to Y_j \text{ in } \hat{\mathcal{E}}^+\}$

16: $\hat{\mathcal{I}}^+ \leftarrow \{(\ell,j) : (\ell,k) \in \hat{\mathcal{I}}^+ \text{ and } (k,j) \in \hat{\mathcal{E}}^+\}$

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

18: **Return:** $\hat{\mathcal{E}}^+$, $\hat{\mathcal{I}}^+$, $\{\widehat{\text{ca}}_G(k)\}_{k=1}^p$

able that empirically depends on the fewest variables in $\mathbf{Y}^-$ as an IV for the working subgraph G^- (line 6), and then chooses the variable in $\mathbf{Y}^-$ that is most strongly correlated with this secondary variable as a leaf node to remove (line 7). The estimated dependence structure is subsequently used to recover the relationships between $\mathbf{Y}_{\text{leaf}(G^-)}$ and $\mathbf{Y} \setminus \mathbf{Y}^-$ (line 11). A rigorous theoretical foundation

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

for these steps is established in Propositions S3.1 and S3.2 of the Supplementary Material. By updating the subgraph G^- and repeating the iterative process, the algorithm ultimately reconstructs the ARG and all candidate IV sets.

4.2 Estimation of causal effects

The estimated ARG obtained from Algorithm 1 provides an initial understanding of the causal structures among the variables in $\mathbf{Y}$. We now show how to estimate the causal effects and hence recover the edges $\hat{\mathcal{E}}$ using the candidate IV sets.

On the basis of the moment condition (3.2), we propose to estimate $\boldsymbol{\beta}^*$ using the generalized method of moments (GMM) (Hansen, 1982). The first step is to construct estimates of the unknown functions $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$ in $\mathbf{M}(\boldsymbol{\beta}^*)$. This can be done in different ways according to the types of candidate IVs. In the case where all variables in $\mathbf{X}_{\mathrm{ca}_G(k)}$ take values in $\{0,1\}$, we follow the idea of Sun et al. (2023) to construct $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$. Let $\alpha(1),\ldots,\alpha(t_k)$ be an enumeration of all subsets $\alpha \subseteq \mathrm{ca}_G(k)$ of cardinality $|\alpha| \geq |\mathrm{ca}_G(k)| - \gamma + 1$. Clearly, for any such α and any subset $\alpha_k \subseteq \mathrm{ca}_G(k)$ with $|\alpha_k| \geq \gamma$, we have $\alpha \cap \alpha_k \neq \emptyset$. This implies that $E\{\Pi_{s\in\alpha}(X_s - \alpha_s) \mid \mathbf{X}_{\mathrm{ca}_G(k)\setminus\alpha_k}\} = 0$, where $\mu_s = E(X_s)$, and hence

$$\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)}) = \left\{\Pi_{s\in\alpha(1)}(X_s - \mu_s), \ldots, \Pi_{s\in\alpha(t_k)}(X_s - \mu_s)\right\}^T.$$

A specific example is given in Example S1.2 of the Supplementary Material. To estimate $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$, one simply substitutes the empirical means $\hat{E}_n(X_s)$ for μ_s

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

in the above expression. If $\mathbf{X}_{\mathrm{ca}_G(k)}$ includes polytomous variables, we break them into dummy variables and compute $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$ similarly. For the continuous case, we use the strategy discussed after Definition 3 to obtain $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$.

Algorithm 2 GMM estimation of $\boldsymbol{\beta}^*$ and $\mathcal{E}$

Input: Data $(\mathbb{X}_{n\times q}, \mathbb{Y}_{n\times p})$, ancestral edge set $\hat{\mathcal{E}}^+$, candidate IV sets $\{\widehat{\mathrm{ca}}_G(k)\}_{k=1}^p$, weighting matrix $\boldsymbol{\Omega}$, FDR level $q^* > 0$

Output: Estimates of $\boldsymbol{\beta}^*$ and $\mathcal{E}$

$N \leftarrow |\hat{\mathcal{E}}^+|$

For each Y_k with descendants in $\hat{\mathcal{E}}^+$, obtain an empirical expression $\hat{\mathbf{Z}}_\gamma(\mathbf{X}_{\widehat{\mathrm{ca}}_G(k)})$ of $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$

For each $(k,j) \in \hat{\mathcal{E}}^+$, $\widehat{\mathrm{me}}_G(k,j) \leftarrow \{i : (k,i) \in \hat{\mathcal{E}}^+, (i,j) \in \hat{\mathcal{E}}^+\}$

For each $(k,j) \in \hat{\mathcal{E}}^+$ and $\boldsymbol{\beta} = (\beta_{kj}) \in \mathbb{R}^N$, obtain an empirical expression of $M_{kj}(\boldsymbol{\beta})$: $\hat{M}_{kj}(\boldsymbol{\beta}) \leftarrow \hat{\mathbf{Z}}_\gamma(\mathbf{X}_{\widehat{\mathrm{ca}}_G(k)})(Y_j - \sum_{i\in\widehat{\mathrm{me}}_G(k,j)} \beta_{ij}Y_i - \beta_{kj}Y_k)$

Concatenate $\hat{M}_{kj}(\boldsymbol{\beta})$ into $\hat{\mathbf{M}}(\boldsymbol{\beta})$ and solve the following problem:

$$\hat{\boldsymbol{\beta}} \leftarrow \arg\min_{\boldsymbol{\beta}} \hat{E}_n\{\hat{\mathbf{M}}(\boldsymbol{\beta})\}^T \boldsymbol{\Omega} \hat{E}_n\{\hat{\mathbf{M}}(\boldsymbol{\beta})\} \tag{4.1}$$

Obtain the standard errors $\hat{\sigma}_{kj}$ of $\hat{\beta}_{kj}$ for all $(k,j) \in \mathcal{E}^+$ by Theorem 3

Calculate the p-values $P_{kj} \leftarrow 2\{1 - \Phi(|\hat{\beta}_{kj}|/\hat{\sigma}_{kj})\}$ for all $(k,j) \in \mathcal{E}^+$

Order the p-values as $P_{(1)} \le \cdots \le P_{(N)}$ with $P_{(i)}$ corresponding to $(k_i, j_i) \in \mathcal{E}^+$

$\ell \leftarrow \max\{i : P_{(i)} \le iq^*/(N\sum_{j=1}^N j^{-1})\}$

$\hat{\mathcal{E}} \leftarrow \{(k_i, j_i)\}_{i=1}^\ell$

Return: $\hat{\boldsymbol{\beta}}$, $\hat{\mathcal{E}}$

After obtaining estimates of $\mathbf{Z}_\gamma(\mathbf{X}_{\mathrm{ca}_G(k)})$, we proceed with the GMM estimation of $\boldsymbol{\beta}^*$ and recovery of $\mathcal{E}$, as summarized in Algorithm 2. Note that the weighting matrix $\boldsymbol{\Omega}$ in (4.1) may affect the asymptotic variance of the GMM estimator. In practice, $\boldsymbol{\Omega}$ can be either specified as the identity matrix or computed from the data as suggested by Hansen (1982). In the next subsection, we derive

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

the asymptotic normality of $\hat{\boldsymbol{\beta}}$ (Theorem 3). The result is used in Algorithm 2 to test whether the individual entries of $\boldsymbol{\beta}^*$ are zero, thereby allowing us to recover the edges in $\mathcal{E}$. To adjust for multiple comparisons, we apply the Benjamini–Yekutieli method (Benjamini and Yekutieli, 2001) in Algorithm 2 to control the false discovery rate (FDR) at level q^*; see Theorem 4.

4.3 Theoretical guarantees

In this subsection, we provide theoretical guarantees for our PLACID method in terms of causal discovery and inference. We begin with the following result, showing that Algorithm 1 consistently learns the ARG and candidate IV sets.

Theorem 2 (Consistency of ancestral structure recovery)**.** *Suppose that Assumptions 1–3 hold and $\alpha = O(n^{-2})$ in Algorithm 1. Then the estimated ARG $\hat{G}^+$ and candidate IV sets $\{\widehat{\mathrm{ca}}_G(k)\}_{k=1}^p$ from Algorithm 1 satisfy*

$$\lim_{n\to\infty} P(\hat{G}^+ = G^+) = 1 \text{ and } \lim_{n\to\infty} P\{\widehat{\mathrm{ca}}_G(k) = \mathrm{ca}_G(k)\} = 1, \quad k = 1, \ldots, p.$$

Theorem 2 relies on the fact that Algorithm 1 with our choice of α can asymptotically detect any dependence between $\mathbf{X}$ and $\mathbf{Y}$ as $n \to \infty$. This property is a consequence of the asymptotic results for DC-based tests (Székely et al., 2007), which are summarized in Section S2 of the Supplementary Material. The significance level α is a hyperparameter similar to that in the PC algorithm (Spirtes et al., 2001). As suggested by Spirtes et al. (2001), we choose $\alpha = O(n^{-2})$ decay-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

ing with n to ensure the convergence to the correct decision with probability one. Assuming consistency of the estimates from Algorithm 1, we have the following result concerning the inference of causal effects and directions by Algorithm 2.

Theorem 3 (Asymptotic normality of $\hat{\boldsymbol{\beta}}$)**.** *Suppose that Assumptions 1–4 hold. Then the estimated causal effects $\hat{\boldsymbol{\beta}}$ from Algorithm 2 satisfy*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

where the asymptotic variance

$$\mathbf{V} = (\mathbf{0}_{q\times q}, \mathbf{I}_{|\mathcal{E}^+|})(\mathbf{G}^T\mathbf{W}_{\boldsymbol{\Omega}}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}_{\boldsymbol{\Omega}}\mathbf{F}\mathbf{W}_{\boldsymbol{\Omega}}\mathbf{G}(\mathbf{G}^T\mathbf{W}_{\boldsymbol{\Omega}}\mathbf{G})^{-1}(\mathbf{0}_{q\times q}, \mathbf{I}_{|\mathcal{E}^+|})^T,$$

and the specific forms of $\mathbf{W}_{\boldsymbol{\Omega}}$, $\mathbf{G}$, and $\mathbf{F}$ are given in Section S3 of the Supplementary Material.

Theorem 3 accounts for the uncertainty due to the unknown mean $\boldsymbol{\mu} = E(\mathbf{X})$ in the construction of surrogate IVs. To this end, we augment the estimating equations for $\boldsymbol{\beta}^*$ with those for $\boldsymbol{\mu}$. The asymptotic variance of $\hat{\boldsymbol{\beta}}$ is then the corresponding submatrix of the usual sandwich estimator for the full asymptotic variance. Additionally, although (2.2) contains nonparametric terms, our estimates still achieve $\sqrt{n}$-consistency. This is not surprising because our surrogate IVs are constructed without approximating the nonparametric terms, thus eliminating the approximation errors inherent in conventional PLM methods. In fact, our estimator is directly derived from the estimating equation (4.1), which

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

contains only parametric components. As a result, $\sqrt{n}$-consistency follows from standard GMM theory.

For an estimated edge set $\hat{\mathcal{E}}$, let TP, RE, and FP denote the numbers of estimated edges with correct directions, those with reverse directions, and those not in the true DAG, respectively. Define the false discovery proportion of $\hat{\mathcal{E}}$ by $\mathrm{FDP}(\hat{\mathcal{E}}) = (\mathrm{RE}+\mathrm{FP})/(\mathrm{TP}+\mathrm{RE}+\mathrm{FP})$, and the false discovery rate of $\hat{\mathcal{E}}$ by $\mathrm{FDR}(\hat{\mathcal{E}}) = E\{\mathrm{FDP}(\hat{\mathcal{E}})\}$. The following result ensures that Algorithm 2 controls the FDR in edge recovery at the nominal level.

Theorem 4 (FDR control in edge recovery)**.** *Suppose that Assumptions 1–4 hold and $\mathcal{E} \neq \emptyset$. Then for any $q^* \in (0,1)$, the estimated edge set $\hat{\mathcal{E}}$ from Algorithm 2 satisfies*

$$\lim_{n\to\infty} \mathrm{FDR}(\hat{\mathcal{E}}) \leq q^*.$$

Both Theorems 3 and 4 require the consistency of the estimated ARG, which is guaranteed by Theorem 2. When the consistency fails to hold, one may consider a post-selection inference framework that projects the true data-generating process onto the selected model (Kuchibhotla et al., 2022; Gradu et al., 2025). However, inference under this framework must be interpreted with caution, as the resulting parameters generally lack a causal interpretation (Berk et al., 2013). In contrast, by leveraging the model selection consistency established in Theorem 2, PLACID provides valid inferences for large samples, while retaining a clear causal

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

interpretation of the parameters.

5. Simulation studies

This section examines the finite-sample performance of our PLACID method. For causal discovery, we compare our method with GrIVET (Chen et al., 2024), RFCI (Colombo et al., 2012), and LRpS-GES (Frot et al., 2019). Since the last two are unable to estimate causal effects, we compare our method only with GrIVET for parameter estimation, where the effects of $\mathbf{X}$ on $\mathbf{Y}$ are specified in a linear form.

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. For the SEM in (2.2), we consider both continuous and discrete cases of secondary variables $\mathbf{X}$. To examine our method for DAGs of different sizes, we fix the sample size at $n = 1000$ while varying the dimensions as $(p, q) = (10, 25)$ and $(20, 50)$. More implementation details are provided in Section S4 of the Supplementary Material.

For causal discovery, RFCI outputs a partial ancestral graph and LRpS-GES outputs a completed partially DAG, both of which may include undirected edges. We evaluate both methods favorably by assuming that the correct directions were

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

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

Graph	p	Method	FDP	TPR	SHD	JI
Random	10	PLACID	0.02(0.08)	0.92(0.21)	0.30(0.67)	0.90(0.22)
		GrIVET	0.57(0.36)	0.49(0.39)	4.06(2.90)	0.27(0.27)
		RFCI	0.07(0.24)	0.48(0.39)	1.31(1.29)	0.47(0.39)
		LRpS-GES	0.70(0.12)	0.97(0.13)	5.04(0.88)	0.30(0.12)
	20	PLACID	0.02(0.07)	0.91(0.15)	0.58(0.98)	0.90(0.16)
		GrIVET	0.83(0.15)	0.43(0.26)	16.38(9.70)	0.14(0.11)
		RFCI	0.03(0.13)	0.59(0.28)	2.07(1.60)	0.59(0.28)
		LRpS-GES	0.70(0.08)	0.99(0.04)	10.55(1.40)	0.30(0.08)
Hub	10	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.40(0.36)	0.48(0.38)	7.55(5.51)	0.41(0.36)
		RFCI	0.01(0.04)	0.58(0.21)	3.78(1.89)	0.58(0.21)
		LRpS-GES	0.39(0.05)	0.83(0.10)	6.30(1.40)	0.55(0.09)
	20	PLACID	0.00(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)
		GrIVET	0.65(0.31)	0.41(0.36)	25.82(13.41)	0.28(0.28)
		RFCI	0.07(0.09)	0.45(0.15)	11.04(3.00)	0.44(0.15)
		LRpS-GES	0.43(0.04)	0.80(0.08)	15.38(2.14)	0.50(0.06)

obtained for undirected edges, as in Li et al. (2024). Four performance metrics for causal discovery are used: false discovery proportion (FDP), true positive rate (TPR), structural Hamming distance (SHD), and Jaccard index (JI). Let TP, RE, and FP be defined as in Section 4.3, and FN the number of missing edges from the true DAG. Then $\mathrm{FDP} = (\mathrm{RE}+\mathrm{FP})/(\mathrm{TP}+\mathrm{RE}+\mathrm{FP})$, $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, $\mathrm{SHD} = \mathrm{FP}+\mathrm{FN}+\mathrm{RE}$, and $\mathrm{JI} = \mathrm{TP}/(\mathrm{TP}+\mathrm{SHD})$. The results for the continuous and discrete cases are summarized in Tables 1 and 2, respectively.

Tables 1 and 2 indicate that PLACID performs best in causal discovery across all scenarios. In particular, it effectively controls the FDP below the nominal level

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

Table 2: Means and standard deviations (in parentheses) of different causal discovery performance metrics for four methods with discrete secondary variables.

Graph	p	Method	FDP	TPR	SHD	JI
Random	10	PLACID	0.01(0.04)	0.92(0.07)	0.07(0.30)	0.91(0.08)
		GrIVET	0.14(0.24)	0.63(0.35)	1.08(1.15)	0.57(0.34)
		RFCI	0.00(0.00)	0.82(0.26)	0.25(0.57)	0.82(0.26)
		LRpS-GES	0.62(0.16)	0.92(0.04)	3.39(1.14)	0.38(0.16)
	20	PLACID	0.03(0.07)	0.98(0.05)	0.33(0.73)	0.95(0.09)
		GrIVET	0.28(0.24)	0.65(0.25)	3.26(2.42)	0.50(0.23)
		RFCI	0.00(0.02)	0.94(0.11)	0.34(0.80)	0.94(0.11)
		LRpS-GES	0.64(0.12)	0.97(0.06)	7.91(1.86)	0.36(0.12)
Hub	10	PLACID	0.00(0.00)	0.96(0.07)	0.39(0.63)	0.96(0.07)
		GrIVET	0.00(0.03)	0.50(0.17)	4.55(1.56)	0.50(0.17)
		RFCI	0.00(0.00)	0.44(0.21)	5.06(1.86)	0.44(0.21)
		LRpS-GES	0.44(0.05)	0.78(0.05)	7.52(1.11)	0.49(0.05)
	20	PLACID	0.00(0.00)	0.97(0.04)	0.54(0.82)	0.97(0.04)
		GrIVET	0.01(0.10)	0.49(0.13)	9.92(3.59)	0.49(0.13)
		RFCI	0.00(0.00)	0.38(0.13)	11.87(2.48)	0.38(0.13)
		LRpS-GES	0.50(0.05)	0.80(0.09)	18.96(3.07)	0.45(0.06)

$q^* = 0.05$, while maintaining a TPR higher than 0.9 for powerful edge detection. As expected, GrIVET struggles with nonlinear relationships between the primary and secondary variables, whereas RFCI and LRpS-GES are less effective in handling large effects of unobserved confounders. In the continuous case, PLACID shows remarkable accuracy for hub graphs, likely because the dependence between $\mathbf{X}$ and $\mathbf{Y}$ is well captured by the empirical DC in these settings. In the discrete case, the performance of RFCI is comparable to that of PLACID for random graphs; however, these metrics are calculated by assuming correct directions for undirected edges in RFCI, giving it an unfair advantage. Moreover, RFCI tends

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

Table 3: Means and standard deviations (in parentheses) of different estimation losses for two methods with continuous and discrete secondary variables.

Setting	Graph	p	Method	L_∞	L_1	L_2
Continuous	Random	10	PLACID	0.23(0.38)	0.35(0.67)	0.26(0.46)
			GrIVET	0.78(0.39)	1.91(1.45)	1.05(0.62)
		20	PLACID	0.37(0.44)	0.68(0.99)	0.45(0.57)
			GrIVET	1.07(0.17)	5.03(1.87)	1.84(0.49)
	Hub	10	PLACID	0.11(0.03)	0.39(0.06)	0.17(0.03)
			GrIVET	0.99(0.29)	7.10(4.03)	2.27(1.00)
		20	PLACID	0.09(0.02)	0.63(0.11)	0.17(0.03)
			GrIVET	1.14(0.13)	18.81(9.02)	3.82(1.27)
Discrete	Random	10	PLACID	0.12(0.21)	0.20(0.45)	0.14(0.28)
			GrIVET	0.76(0.38)	1.47(1.08)	0.95(0.55)
		20	PLACID	0.21(0.36)	0.40(0.71)	0.26(0.45)
			GrIVET	0.98(0.20)	3.26(1.60)	1.56(0.50)
	Hub	10	PLACID	0.38(0.35)	0.93(0.57)	0.48(0.37)
			GrIVET	1.16(0.03)	9.00(0.34)	3.02(0.11)
		20	PLACID	0.36(0.22)	2.31(1.35)	0.95(0.60)
			GrIVET	1.18(0.02)	19.99(0.53)	4.50(0.12)

to be less powerful for hub graphs, while PLACID consistently exhibits superior and stable performance across different graph structures and sizes.

For parameter estimation, we compare our method with GrIVET in terms of entrywise L_∞, L_1, and L_2 losses, as reported in Table 3. These results demonstrate the superior performance of PLACID over GrIVET in parameter estimation across various settings. It is also interesting to note that the performance of PLACID in causal discovery and parameter estimation exhibits the same trend, as can be seen from a comparison of Tables 1 and 2 with Table 3. This is reasonable since a better estimate of the ARG enables more accurate parameter estimation, which

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

in turn leads to a more precise recovery of causal structures.

To empirically assess the strength of the surrogate IVs used in PLACID, we follow Stock et al. (2002) and adopt the first-stage F-statistic from two-stage least squares. Specifically, for each primary variable Y_k and its surrogate IVs, we compute the corresponding first-stage F-statistic and then report the average across all primary variables. The resulting average first-stage F-statistics are presented in Table S4.1 of the Supplementary Material, where 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. In Section S5 of the Supplementary Material, we further conduct simulations to more comprehensively evaluate the performance of PLACID under different settings, including varying IV strengths and sample sizes. Section S5 also includes simulations that empirically examine the roles of Assumptions 1 and 3.

6. Application to ADNI data

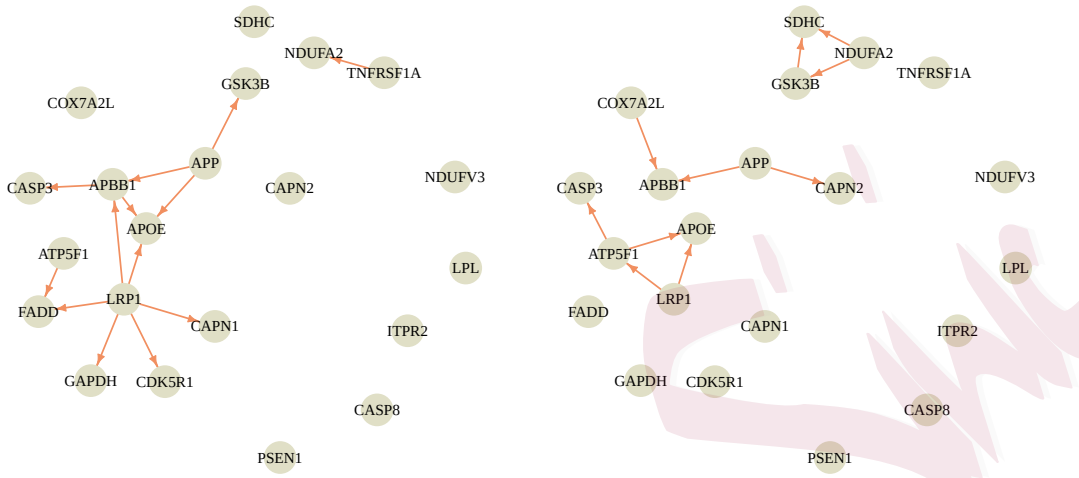
Inferring gene regulatory networks is crucial for understanding the pathophysiology of complex diseases and developing effective therapeutics (Barabási et al., 2011). In this section, we apply our method to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (https://adni.loni.usc.edu) for estimating gene regulatory networks. We use the preprocessed data from Chen et al. (2024),

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

with gene expression levels normalized and adjusted for baseline covariates. By selecting genes with at least one strongly associated single nucleotide polymorphism (SNP) and two strongest SNPs for each gene, the dataset includes $p = 21$ genes as primary variables and $q = 42$ SNPs as secondary variables. Participants were divided into 462 cases with Alzheimer's disease or mild cognitive impairment (AD-MCI) and 247 cognitively normal controls (CN). Partial residual plots in Section S6 of the Supplementary Material suggest nonlinear relationships between some primary and secondary variables, and hence model (2.2) is appropriate. We then apply PLACID to learn the DAGs among the genes for both groups. Chen et al. (2024) assumed that the candidate IV set for each primary variable satisfies the majority rule, namely that more than half of the relevant IVs are valid. Here, we adopt a more conservative choice of $\gamma = 1$, which requires only at least one valid IV per primary variable and is sufficient for PLACID to be applicable. To assess whether Assumption 4 holds with this choice of γ, we empirically evaluate the surrogate IV strength. The average first-stage F-statistics for the AD-MCI and CN groups are 309.96 and 160.06, respectively, both well above the conventional threshold of 10 for weak IVs (Staiger and Stock, 1997), confirming that the selected surrogate IVs are sufficiently strong.

The estimated DAGs for the AD-MCI and CN groups are displayed in Figure 3, which reveal both common and distinctive features of gene regulatory inter-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331



(a) AD-MCI (b) CN

Figure 3: Estimated gene regulatory networks for (a) AD-MCI and (b) CN groups.

actions in the two groups. Compared with the CN group, the AD-MCI group has substantially more edges originating from LRP1, suggesting a critical role of LRP1 in the pathogenesis of Alzheimer's disease. Indeed, it has been known that LRP1 is a major regulator of amyloid-β and tau, the two hallmark proteins in Alzheimer's disease (Bloom, 2014), and contributes to their accumulation and spread in the brain (Rauch et al., 2020). Among the outgoing edges of LRP1, the link to APOE is shared by both groups, which is consistent with the previous finding that LRP1 regulates brain APOE and cholesterol metabolism (Liu et al., 2007). In fact, APOE has long been established as the strongest genetic risk factor for late-onset Alzheimer's disease, and has multifaceted effects on many neurobi-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

ological processes underlying Alzheimer's disease (Serrano-Pozo et al., 2021). We further note that APP is connected to different downstream genes between the two groups. In particular, the AD-MCI group includes paths from APP to APOE, GSK3B, and APBB1, whereas the first two paths are absent in the CN group. APP is the precursor to amyloid-β, whose abnormal processing has been found central to the development of Alzheimer's disease (O'Brien and Wong, 2011). Interestingly, the paths from APP to APOE support the possibility that APOE and cholesterol levels are modulated, directly or indirectly, by APP (Liu et al., 2007).

7. Discussion

We have proposed a novel method for identifying and inferring DAGs under unobserved confounding using invalid IVs. Our method may suffer from certain limitations and can be extended in several directions. First, Assumption 1 may be relaxed to allow dependence among secondary variables. To block paths through correlated secondary variables, one can apply the notion of conditional distance correlation (Wang et al., 2015) to test the independence of X_i and Y_j conditional on the other secondary variables. For parameter estimation, one may follow Sun et al. (2023) and adjust moment conditions with weights accounting for dependence. Second, it would be valuable to extend our setting to nonlinear causal models with unobserved confounders, as in Agrawal et al. (2023). Finally, it

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

would be worthwhile to extend our method to high-dimensional settings where p or q is large. One possible strategy for such extensions is to first estimate the ARG via a DC-based feature screening procedure to assess the dependence between $\mathbf{X}$ and $\mathbf{Y}$ (Li et al., 2012), followed by a high-dimensional GMM estimator (Caner, 2009) to recover the causal effects and directions. We leave these topics for future research.

Supplementary Material

The Supplementary Material includes examples, proofs of the theoretical results, details on simulation settings, additional simulation studies, and additional analysis results for the application.

Acknowledgments

We sincerely thank the editor, associate editor, and two reviewers for their valuable comments, which led to a significant improvement of our paper. Zou and Lin's research was supported by the National Natural Science Foundation of China (12171012, 12292980, and 12292981). Li's research was supported by the National Natural Science Foundation of China (12471269) and National Key R&D Program of China (2022YFA1008100). The public computing cloud from Renmin University of China was used to perform the simulation and data analysis.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

References

Agrawal, R., C. Squires, N. Prasad, and C. Uhler (2023). The DeCAMFounder: Nonlinear causal discovery in the presence of hidden variables. *Journal of the Royal Statistical Society, Series B 85*(5), 1639–1658.

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

Barabási, A.-L., N. Gulbahce, and J. Loscalzo (2011). Network medicine: A network-based approach to human disease. *Nature Reviews Genetics 12*(1), 56–68.

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.

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

Bloom, G. S. (2014). Amyloid-β and tau: The trigger and bullet in Alzheimer disease pathogenesis. *JAMA Neurology 71*(4), 505–508.

Bowden, J., G. Davey Smith, and S. Burgess (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology 44*(2), 512–525.

Caner, M. (2009). Lasso-type GMM estimator. *Econometric Theory 25*(1), 270–290.

Castro, D. C., I. Walker, and B. Glocker (2020). Causality matters in medical imaging. *Nature*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

Communications 11(1), 3673.

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.

Chen, S., Z. Lin, X. Shen, L. Li, and W. Pan (2023). Inference of causal metabolite networks in the presence of invalid instrumental variables with GWAS summary data. *Genetic Epidemiology 47*(8), 585–599.

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

Dominici, F., A. McDermott, and T. J. Hastie (2004). Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association 99*(468), 938–948.

Engle, R. F., C. W. J. Granger, J. Rice, and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association 81*(394), 310–320.

Florens, J.-P., J. Johannes, and S. Van Bellegem (2012). Instrumental regression in partially linear models. *The Econometrics Journal 15*(2), 304–324.

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

Gradu, P., T. Zrnic, Y. Wang, and M. I. Jordan (2025). Valid inference after causal discovery. *Journal of the American Statistical Association 120*(550), 1127–1138.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

REFERENCES

Guo, Z., H. Kang, T. T. Cai, and D. S. Small (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 80*(4), 793–815.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Härdle, W., H. Liang, and J. Gao (2000). *Partially Linear Models*. Berlin: Springer.

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

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

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics 33*(4), 474–484.

Kuchibhotla, A. K., J. E. Kolassa, and T. A. Kuffner (2022). Post-selection inference. *Annual Review of Statistics and Its Application 9*, 505–527.

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

Li, C., X. Shen, and W. Pan (2024). Nonlinear causal discovery with confounders. *Journal of the American Statistical Association 119*(546), 1205–1214.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Li, W., R. Duan, and S. Li (2024). Discovery and inference of possibly bi-directional causal relationships with invalid instrumental variables. *arXiv preprint arXiv:2407.11646*.

Liang, H., S. Wang, J. M. Robins, and R. J. Carroll (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association 99*(466), 357–367.

Liu, Q., C. V. Zerbinatti, J. Zhang, H.-S. Hoe, B. Wang, S. L. Cole et al. (2007). Amyloid precursor protein regulates brain apolipoprotein E and cholesterol metabolism through lipoprotein receptor LRP1. *Neuron 56*(1), 66–78.

Neto, E. C., M. P. Keller, A. D. Attie, and B. S. Yandell (2010). Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics 4*(1), 320–339.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica 58*(4), 809–837.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In *Econometrics*, Volume 11 of *Handbook of Statistics*, pp. 419–454. Amsterdam: North-Holland.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Oates, C. J., J. Q. Smith, and S. Mukherjee (2016). Estimating causal structure using conditional DAG

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

models. *Journal of Machine Learning Research 17*(54), 1–23.

O'Brien, R. J. and P. C. Wong (2011). Amyloid precursor protein processing and Alzheimer's disease. *Annual Review of Neuroscience 34*, 185–204.

Ongen, H., A. A. Brown, O. Delaneau, N. I. Panousis, A. C. Nica, G. Consortium et al. (2017). Estimating the causal tissues for complex traits and diseases. *Nature Genetics 49*(12), 1676–1683.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics 81*(3), 559–575.

Rauch, J. N., G. Luna, E. Guzman, M. Audouard, C. Challis, Y. E. Sibih et al. (2020). LRP1 is a master regulator of tau uptake and spread. *Nature 580*(7803), 381–385.

Robins, J. M., S. D. Mark, and W. K. Newey (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics 48*(2), 479–495.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

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

Runge, J., P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances 5*(11), eaau4996.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

REFERENCES

Serrano-Pozo, A., S. Das, and B. T. Hyman (2021). APOE and Alzheimer's disease: Advances in genetics, pathophysiology, and therapeutic approaches. *The Lancet Neurology 20*(1), 68–80.

Spirtes, P., C. Glymour, and R. Scheines (2001). *Causation, Prediction, and Search* (2nd ed.). Cambridge, MA: MIT Press.

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

Stock, J. H., J. H. Wright, and M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics 20*(4), 518–529.

Sun, B., Z. Liu, and E. J. Tchetgen Tchetgen (2023). Semiparametric efficient G-estimation with invalid instrumental variables. *Biometrika 110*(4), 953–971.

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.

Tchetgen Tchetgen, E. J., J. M. Robins, and A. Rotnitzky (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika 97*(1), 171–180.

Triantafillou, S., V. Lagani, C. Heinze-Deml, A. Schmidt, J. Tegner, and I. Tsamardinos (2017). Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells. *Scientific Reports 7*, 12724.

Wang, X., W. Pan, W. Hu, Y. Tian, and H. Zhang (2015). Conditional distance correlation. *Journal of the American Statistical Association 110*(512), 1726–1734.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0331

REFERENCES

Windmeijer, F., H. Farbmacher, N. Davies, and G. Davey Smith (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.

Ye, T., J. Shao, and H. Kang (2021). Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *The Annals of Statistics 49*(4), 2079–2100.

Zhao, Q., J. Wang, G. Hemani, J. Bowden, and D. S. Small (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics 48*(3), 1742–1769.

Zilinskas, R., C. Li, X. Shen, W. Pan, and T. Yang (2024). Inferring a directed acyclic graph of phenotypes from GWAS summary statistics. *Biometrics 80*(1), ujad039.

School of Mathematical Sciences, Peking University, Beijing, China, 100871. E-mail: (jingzou@stu.pku.edu.cn)

Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China, 100872. E-mail: (weilistat@ruc.edu.cn)

School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China, 100871. E-mail: (weilin@math.pku.edu.cn)